DOCUMENT RESUME

ED 117 770	CS 501 221
TITLE	Status Report on Speech Research: A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for Its Investigation, and Practical Applications, April 1 - September 30, 1975.
INSTITUTION	Haskins Labs., New Haven, Conn.
REPORT NO	SR-42/43(1975)
PUB DATE	Nov 75
NOTE	304p.
EDRS PRICE	MF-$0.83 HC-$16.73 Plus Postage
DESCRIPTORS	*Articulation (Speech); *Auditory Perception; Educational Research; Higher Education; Perception; Perceptual Development; Reports; Research; Research Reviews (Publications); *Speech; Visual Perception; Vowels

ABSTRACT

This status report on speech research includes 16 essays and extended reports. Included are "Perspectives in Vision: Conception of Perception?" "The Perception of Speech," "The Dynamic Use of Prosody in Speech Perception," "Speech and the Problem of Perceptual Constancy," "Coperception," "Dichotic 'Masking' of Voice Onset Time," "The Number Two and the Natural Categories of Speech and Music," "Processing Two Dimensions of Nonspeech Stimuli," "Predicting Initial Cluster Frequencies by Phonemic Difference," "Hemispheric Specialization for Speech Perception in Four-Year Old Children from Low and Middle Socioeconomic Classes," "Automatic Segmentation of Speech into Syllabic Units," "Pushing the Voice Onset Time Boundary," "Some Maskinglike Phenomena in Speech Preception," "The Preception of Vowel Duration in Vowel-Consonant and Consonant-Vowel-Consonant Syllables," "Accounting for the Poor Recognition of Isolated Vowels," and "Some Acoustic Measures of Anticipatory and Carryover Coarticulation." (TS)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED117770

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

SR-42/43 (1975)

Status Report on

SPEECH RESEARCH

A Report on
the Status and Progress of Studies on
the Nature of Speech, Instrumentation
for its Investigation, and Practical
Applications

1 April - 30 September 1975

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports.)

CS 501 221

ERRATA

Personnel list: For Michael T. Truvey read Michael T. Turvey

Susan C. Polgar should not be on the list

Page 46, paragraph 3, line 14: For silhoustte read silhouette

Page 189, paragraph 2, lines 6-7: For condition read cognition

ACKNOWLEDGMENTS

The research reported here was made possible in part by support from the following sources:

National Institute of Dental Research
Grant DE-01774

National Institute of Child Health and Human Development
Grant HD-01994

Research and Development Division of the Prosthetic and Sensory Aids Service, Veterans Administration
Contract V101(134)P-342

Advanced Research Projects Agency, Information Processing Technology Office, under contract with the Office of Naval Research, Information Systems Branch
Contract N00014-67-A-0129-0002

United States Army Electronics Command, Department of Defense
Contract DAAB03-75-C-0419(L 433)

National Institute of Child Health and Human Development
Contract NIH-71-2420

National Institutes of Health
General Research Support Grant RR-5596

HASKINS LABORATORIES

Personnel in Speech Research

Alvin M. Liberman,* President and Research Director
Franklin S. Cooper, Associate Research Director
Patrick W. Nye, Associate Research Director
Raymond C. Huey, Treasurer
Alice Dadourian, Secretary

Investigators

Arthur S. Abramson*
Fredericka Bell-Berti*
Gloria J. Borden*
Earl Butterfield[1]
James E. Cutting*
Christopher J. Darwin[2]
Ruth S. Day*
Michael F. Dorman*
Peter Eimas[3]
Jane H. Gaitenby
Thomas J. Gay*
Katherine S. Harris*
Leigh Lisker*
Ignatius G. Mattingly*
Paul Mermelstein
Seiji Niimi[4]
Lawrence J. Raphael*
Donald P. Shankweiler*
George N. Sholes
Michael Studdert-Kennedy*
Michael T. Truvey*
Tatsujiro Ushijima[4]

Technical and Support Staff

Eric L. Andreasson
Dorie Baker*
Elizabeth P. Clark
Cecilia C. Dewey
Janneane F. Gent
Donald S. Hailey
Harriet G. Kass*
Diane Kewley-Port*
Sabina D. Koroluk
Christina R. LaColla
Roderick M. McGuire
Agnes McKeon
Terry F. Montlick
Susan C. Polgar*
Loretta J. Reiss
William P. Scully
Richard S. Sharkany
Edward R. Wiley
David Zeichner

Students*

Mark J. Blechner
Susan Brady
John Collins
David Dechovitz
Susan Lea Donald
G. Campbell Ellison
Donna Erickson
F. William Fischer
Carol A. Fowler
Frances J. Freeman
Gary M. Kuhn
Andrea G. Levitt
Roland Mandler
Robert F. Port
Robert Remez
Philip E. Rubin
Helen Simon
James M. Vigorito

---

*Part-time
[1] Visiting from the University of Kansas, Lawrence.
[2] Visiting from the University of Sussex, Brighton, England.
[3] Visiting from Brown University, Providence, R. I.
[4] Visiting from University of Tokyo, Japan.

CONTENTS

I. Manuscripts and Extended Reports

Perspectives in Vision: Conception or Perception? -- M. T. Turvey. 1

The Perception of Speech -- C. J. Darwin. 59

On the Dynamic Use of Prosody in Speech Perception -- C. J. Darwin. 103

Speech and the Problem of Perceptual Constancy -- Donald Shankweiler, Winifred Strange, and Robert Verbrugge. 117

"Coperception": A Preliminary Study -- Bruno H. Repp 147

Dichotic "Masking" of Voice Onset Time -- Bruno H. Repp 159

The Magical Number Two and the Natural Categories of Speech and Music -- James E. Cutting. 189

Processing Two Dimensions of Nonspeech Stimuli: The Auditory-Phonetic Distinction Reconsidered -- Mark J. Blechner, Ruth S. Day, and James E. Cutting. 221

Predicting Initial Cluster Frequencies by Phonemic Difference -- James E. Cutting. 233

Hemispheric Specialization for Speech Perception in Four-Year-Old Children from Low and Middle Socioeconomic Classes -- Donna S. Geffner and M. F. Dorman. 241

Automatic Segmentation of Speech into Syllabic Units -- Paul Mermelstein. . 247

On Pushing the Voice-Onset-Time Boundary About -- Leigh Lisker, Alvin M. Liberman, Donna Erickson, and David Dechovitz. 257

Some Maskinglike Phenomena in Speech Perception -- M. F. Dorman, L. J. Raphael, A. M. Liberman, and B. Repp. 265

The Perception of Vowel Duration in VC and CVC Syllables -- Lawrence J. Raphael, Michael F. Dorman, and A. M. Liberman. 277

On Accounting for the Poor Recognition of Isolated Vowels -- Donald Shankweiler, Winifred Strange, and Robert Verbrugge. 285

Some Acoustic Measures of Anticipatory and Carryover Coarticulation -- Fredericka Bell-Berti and Katherine S. Harris 297

II. Publications and Reports. 305

III. Appendix: DDC and ERIC numbers (SR-21/22 - SR-41). 307

I. MANUSCRIPTS AND EXTENDED REPORTS

Perspectives in Vision: Conception or Perception?*

M. T. Turvey[+]

ABSTRACT

To a very large extent theories of perception assume that the stimulus at the receptors underdetermines the perceptual experience. Consequently, the official doctrine holds that perception is predicated on conception: one must exploit one's knowledge about the world in order to perceive it. The manifestation of this point of view in the currently popular information-processing approach to visual perception is treated at some length. An alternative approach begins with the argument that the correspondence between the ambient optic array and the environment, the "what" of visual perception, has been insufficiently examined and that the enterprise of modeling perceptual processes, the "how" of perception, is premature. Indeed, the alternative position suspects that stimulation does not underdetermine perception and that, contrary to tradition, perception is not mediated by intellectual processes. This thesis of direct perception is developed and contrasted with that of indirect perception.

From what cloth shall we cut the theory of visual perception? For most scholars over the centuries the answer has been singularly straightforward: conception. The official doctrine is that our visual perception of the world depends in very large part on our conception of it. To paraphrase: knowing the world perceptually rests on knowing about the world conceptually. The inquiry into visual perception is dominated by this thesis of conception as primary, and the first charge of this paper is to examine in elementary but reasonably detailed fashion the manifestations of this thesis in current theory and research.

But there is another and contrasting point of view to which this paper will turn in due course; one that asserts the primacy of perception. From this standpoint, information about the visual world is obtained without the

---

*This paper was presented as an invited State-of-the-Art address to the World Congress on Dyslexia sponsored by the Orton Society and Mayo Clinic, Rochester, Minn., 23-25 November 1974. It is to be published in Reading, Perception and Language, ed. by M. Rawson and D. Duane. (Baltimore, Md.: York).

[+] Also University of Connecticut, Storrs.

Acknowledgment: The preparation of this manuscript was supported in part by a John Simon Guggenheim Fellowship awarded to the author for the period 1973-1974.

intervention of conceptual processes. On this view, conceptual knowledge is the offspring of perception. To declare the independence of perception from conception is to declare the dependence of perception on stimulation. The second charge of this paper is to examine the meaning and implications of these declarations.

INDIRECT PERCEPTION: CONSTRUCTIVISM AND THE PRIMACY OF CONCEPTION

Introduction

Only relatively superficial differences divide the various theories of visual perception with which we are most acquainted. With respect to the basic postulates on which each is founded there is an overwhelming consistency of opinion: they all concur to a greater or lesser degree that perception begins with retinal data that relate imperfectly, even ambiguously, to their source--the world of objects, surfaces, and events.

Consequently, it is customary to liken the task of the perceiver to that of a detective who must seek to determine what transpired from the bits and pieces of evidence available to him. Extending this metaphor, we have supposed that there are "cues" or "clues" to be read from the retinal image, and to a very large extent the endeavors of students of perception have been directed to isolating these cues and inquiring about how they are used. The impression is that a large discrepancy exists between what is given in the retinal image and what is perceived; like Gregory (1969), we wonder how so little information can control so much behavior. From necessity we argue that the perceiver must engage in a large stock of inferential and hypothesis-generating and testing procedures that rely heavily on memory--on internal models of reality established in the course of prior experience of both the individual and the species.

This general approach to the theory of visual perception has a long tradition. Recall that John Locke drew a distinction between primary qualities as given and secondary qualities as inferred, and that Helmholtz coined the phrase "unconscious inference" to represent this theoretical persuasion. Currently it is given expression through the term "constructivism," which will be used in this paper for any theory proposing that in order to perceive one must go beyond what is given in stimulation. Thus, we understand constructivism to mean that perception cannot be achieved directly from stimulation (one might say that stimulation underdetermines perceptual experience). On the contrary, perceptual experience is constructed or created out of a number of ingredients, only some of which are provided by the data of the senses. Other ingredients in a perception recipe are provided by our expectations, our biases, and primarily by our conceptual knowledge about the world. The gist of the constructive interpretation is conveyed in the following remark: "...perceptions are constructed by complex brain processes from fleeting fragmentary scraps of data signalled by the senses and drawn from the brain's memory banks--themselves constructions from snippets of the past"(Gregory, 1972:707). Helmholtz (1925), in response to the ambiguity of retinal stimulation, might have said: we perceive that object or that event that would normally fit the proximal stimulus distribution (Hochberg, 1974).

Internal Modeling of the Relation Structure of External Events

The view of the brain as a complex information-processing mechanism that internally models the world is of central importance to theories of the constructivist persuasion. This view, represented best by Craik (1943), can be stated explicitly: neural processes or mental operations symbolically mirror objects, events, and their interrelationships.

The advantage of internal modeling is that outcomes can be predicted. Mental processes allow us to proceed vicariously through a pattern of motions or a succession of events much as an engineer determines a reliable design for a bridge *before* he begins building. In short, we can try out mentally what would occur without actually performing the test, the outcome of which may be useless or even harmful. There would seem to be little reason to debate this property of mind with respect to thought and language, but can we with the same confidence regard visual perception analogously as a modeling, imitative, predictive process? The constructivist answer is an unequivocal, "Yes." We may conceptualize visual perception as the task of creating a short-term model of contemporary distal stimuli out of, on the one hand, the contemporary but crude proximal stimulation, and on the other, from the internalized long-term model of the world (cf. Arbib, 1972). This is a significant feature of constructivism: it encourages us to examine the possibility that psychological processes such as thought, language, and seeing are more similar than they are different. The kinds of knowledge--heuristics, algorithms, and so on--that permit thought may not differ significantly from those that permit language, and these in turn may be equivalent to those that yield visual perception. Paraphrasing Kolers (1968), Katz (1971), Sutherland (1973), and others, there is nothing that would suggest to a constructivist different kinds of intelligence underlying these apparently different activities.

That brain mechanisms can model the world in the sense of exhibiting processes that have a similar relational structure to sets of physical events, is an intuition that is currently receiving some measure of support in the laboratory. Echoing Craik's (1943) hypothesis, Shepard (in press; Shepard and Chipman, 1970) proposes a second-order isomorphism between physical objects and their internal representations. This isomorphism is not in the first-order relation between a particular object and its internal representation (as Gestalt psychologists used to have it), but rather in the relation between (1) the relations among a set of objects and (2) the relations among their corresponding internal representations. To quote Shepard and Chipman (1970), "Thus, although the internal representation of a square need not itself be a square it should (whatever it is) at least have a closer functional relation to the internal representation of a rectangle than to, say, a green flash or a taste of persimmon."

There are two experiments that speak favorably for the notion of a second-order isomorphism. In the first (Shepard and Chipman, 1970), fifteen states of the United States--ones that did not differ too greatly in size--were selected for similarity judgments. One hundred and five pairings of members of this set were presented in name form and ranked by the subjects according to similarity. The same subjects then ranked the same 105 pairs of states presented in picture form. The structure of similarity relations among the shapes was virtually identical whether the states compared were there to be perceived (pictorial presentation) or could only be imagined (name presentation). Moreover, for both name and pictorial presentation, the similarity judgments corresponded to identifiable properties of the actual cartographic shapes of the states (see also Gordon and Hayward, 1973).

In the second experiment (Heider and Oliver, 1972), a second-order isomorphism is implied between the structure of color space as represented in memory and the structure of color space as given in perceptual experience. They sought to determine whether two people who differed markedly in color terminology (the Dani of New Guinea and Americans) actually structured the color space in markedly different ways. There were two tasks: in one, subjects from the two populations named colors; in the other, they matched colors from memory. Through multidimensional scaling, it was determined that the structure of the color space was remarkably similar for the two populations when derived from the memory data but, as expected, quite dissimilar when derived from naming. The important comparison here lies between the relational structure of colors in memory, and the relational structure of colors in perception (Shepard, 1962; Helm, 1964). On the available evidence, perceptually the two structures appear to be virtually the same (Heider and Oliver, 1972).

One may argue from these experiments that memory preserves or mimics the structural relations among perceptual properties. But the notion of brain mechanisms modeling external events suggests something more: an isomorphism between external and internal processes. We turn, therefore, to experiments that examine the following proposition as a corollary of the second-order isomorphism theorem: when one is imaging an external process, one passes through an orderly set of internal states related in a way that mimics the relations among the successive states of the external process (cf. Shepard and Feng, 1972).

Suppose that you are shown a pair of differently oriented objects pictured in Figure 1 and that you have to determine whether the two objects are the same or different. I suspect that you would reach your decision by manipulating the objects in some way, say, by rotating one of them and comparing perspectives. But suppose that you are shown a pair of two-dimensional portrayals of the three-dimensional objects (which, of course, is what Figure 1 is) and that your task is to decide as quickly as possible (and obviously without the aid of manipulation) whether one of the objects so depicted could be rotated into the other. In an experiment of this kind (Shepard and Metzler, 1971), it was shown that the decision latency was an increasing linear function of the angular difference in the portrayed orientation of the two objects. At 0° difference, the latency was 1 sec, while at 180° difference the latency was 4 or 5 sec. Each additional degree of rotation added approximately 16 msec to the latency of recognition. This was essentially so whether the rotation was in the plane of the picture or in depth.

This example illustrates the capability of neural processes to model the relational structure of external happenings. For further instances, the reader is referred to Cooper and Shepard (1973) and Shepard and Feng (1972).

What Presupposes Indirect Perception?

What is behind the assumption that stimulation underdetermines perception and that to perceive visually one must go beyond what is given in the light to an ocular system? What is the given? History's answer is that the reference is to the stimuli for the receptors or to the sensations provided by the senses. Clearly, the need to suggest that a brain must guess, infer, or construct will have its roots primarily in the attributes that tradition has adduced for stimuli and/or sensations.

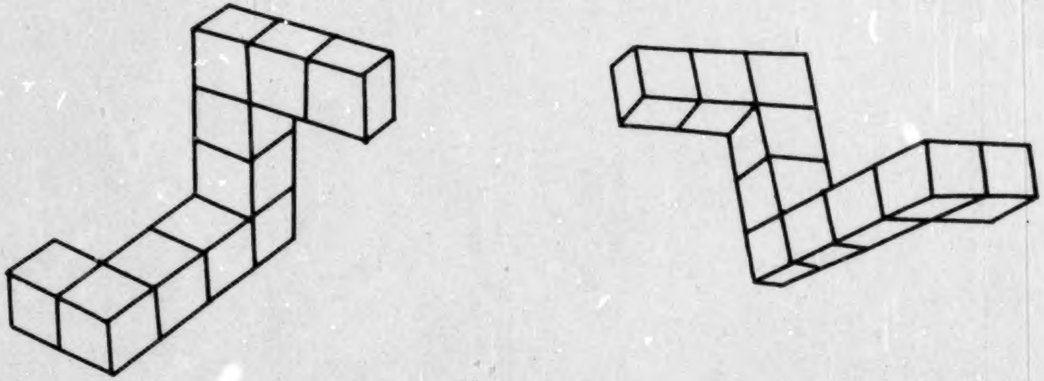

Figure 1: Are these two perspectives of the same object? (After Shepard and Metzler, 1971.)

FIGURE 1

We can begin with the notion that they are given and given directly. Construction presupposes some basic materials, and therefore all constructive theories concede a preliminary stage whose output is not guessed or created. And the degree to which a theory conceives perception as indirect is inversely related to the elaborateness of the evidence deemed to be directly detectable. The point sensations that occupied this preliminary stage in the thinking of the structuralists have given way to the oriented lines and angles imposed upon us by an impressive neurophysiology (Hubel and Wiesel, 1967, 1968) and the pragmatics of building workable computer programs for pattern recognition (Minsky, 1963).

A current idea is that preliminary to object recognition is a stage that detects primitive features. It is assumed as a matter of principle that there is a finite number of such features, which poses the problem of how an indefinitely large number of objects can be discriminated and recognized. In order that infinite usage may be made of a limited number of features, subsequent processes in perception must instantiate knowledge about relations among features and must, perhaps, be capable of generating a variety of object representations (Minsky, 1963). As one theorist (Gregory, 1970:36) has remarked: "Perception must, it seems, be a matter of seeing the present with stored objects from the past."

The assumption that what is given directly in perception is a finite set of punctate elements necessitates a search beyond the stimulus for an explanation of the perception of objects and of the fact that we experience optical events as spatially unified. The assumption of stimuli as punctate demands a theory of indirect, constructed perception.

It is curious that the the circumscribed feature set, so far uncovered by neurophysiology, has been adopted energetically and somewhat uncritically as the departure point for much of current visual perception theory. The experimental evidence is that there are neural units selectively sensitive to simple spatial relations (e.g., a line in a given orientation) and to simple spatial-temporal relations (e.g., a line moving in a given direction). Does this evidence delimit the list of feature detectors? Clearly, there is no a priori reason for believing that future research will not reveal detectors, or more aptly, systems, selectively sensitive to more complex optical relations. The growing evidence for selective sensitivity to spatial frequency (Sekuler, 1974) corroborates this suspicion.

But whatever our misgivings, we should not lose sight of the important and instructive fact that the above "simple" spatial and spatial-temporal relations are characterized as being directly given. In using the term "direct" with respect to the detection of features, we intend the following. First, the course of detection does not involve establishing an internal representation of the feature that is then matched by some separate process against a stored representation. Second, and related, the pickup of these features does not depend on information other than that currently available in the stimulation; that is to say, their detection is not underdetermined.

Closely allied with the interpretation of stimuli as punctate elements, is the prevailing assertion that stimuli are temporally discrete. If stimuli are momentary, then there must be routines for combining stimuli that are distributed in time. This is exemplified in the familiar problem of how scenes are perceived through a succession of eye movements. An aim of the fovea corresponds to the reception of a sample of the available stimuli. Thus a succession of aims yields a succession of retinal images and therefore a succession of different sets of punctate elements. For the observer who moves or scans, the fitting-together operations performed on each set of adjacent punctate elements must be supplemented by processes that collect together and synthesize the products of these operations over time.

Hochberg (1968, 1970) and Neisser (1967) have emphasized the distinction between the information in a single glance and the integration of information from a succession of glances. They argue that a panoramic impression cannot be specified by a single fixation, for in general a single fixation provides the observer with only local information about the three-dimensional structure of a scene. Moreover, only in the foveal region of a fixation is the information detailed. Thus, to perceive a scene, an observer must integrate several fixations. Evidently panoramic perception is constructed, but how this construction-through-integration takes place remains very much a mystery and to date there are no viable accounts of how it might occur.

The characterization of stimuli as punctate and momentary supplies the backdrop for one of the more heralded reasons for claiming that perception goes beyond what is given. The well-known phi phenomenon provides a case in point. If two lamps are lit in alternation, with an appropriate interval elapsing between successive lightings, one perceives a single lamp moving back and forth. Punctate and momentary characterization of stimuli would indicate that there are actually two stimuli separate in space and time, although perceptually the impression is of a single stimulus moving from one point in space to another. Thus, we appear to have a compelling reason for the notion that perception goes beyond what is given: perceiving a stimulus where it is not.

One particularly dramatic variant of the phi phenomenon, which would seem to dictate rather than invite the constructive interpretation, is provided by Kolers (1964). When a Necker cube is set into oscillating apparent motion, an observer under the appropriate conditions will see the cube as rotating in "mid-flight." In similar circumstances, one can also see circles elastically transformed into squares and upward pointing arrows rigidly transformed into downward pointing arrows (Kolers and Pomerantz, 1971). We see a changing form where there is no form changing. Moreover, the particular transformation experienced befits the forms involved. Based on these observations, seeing mirrors thinking.

If perception is indirect, we may ask: What do perception by eye and perception by hand have in common? We might answer "very little" on the assumption that an eye delivers to a brain sensations or features that are radically different from those that a hand delivers. So how is it that I can often perceive by eye that which I can also perceive by touch? Traditionally, the answer to this question has been sought in a mediating link--frequently association--which relates visual impressions to tactile impressions, and vice versa. In this view, the data of one sense are rationalized--given meaning--by the data of another.

Where we conceptualize the senses as yielding distinctively different data, we are forced into assuming that cross-modality correlation is an essential feature of perception. This is especially so where we believe that the data of one sense are intrinsically less meaningful than the data of another. Since the time of Bishop Berkeley, vision has been construed, repeatedly, as parasitic on touch and muscle kinesthesis.

Let us now turn to pictures, for they would appear to provide prima facie evidence that perception must go beyond what is given in the stimulation. Pictures are flat projections of three-dimensional configurations and they can be visualized in either way. Further, it is quite obvious that we can "see" the three-dimensional structure represented by a picutre even when the information is peculiarly impoverished, as witnessed by outline drawings. We can also "see" the appropriate three-dimensional arrangement, even though the picture is ambiguous in the sense that the same projected form could arise from an infinite variety of shapes. In Goodman's (1968) view, pictorial structure is an arbitrary conventional language that must be learned in a way that corresponds to how we learn to read. For Hochberg (1968) the resolution of the problem of picture processing is sought in the schematic maps stored in one's memory banks and evoked by features. Similarly, Gregory (1970) has looked to object-hypotheses as the tools by which we disambiguate a two-dimensional portrayal of a three-dimensional scene. Though none of these authors provides anything like an account of how "other knowledge" is brought to bear on picture processing, the task has been taken up in earnest by workers in artificial intelligence. It is worth noting that there are students of perception who, contrary to the above points of view, suspect that the perception of pictures need not be indirect (Gibson, 1971; Hagen, 1974).

Finally, there is the idea that the environment is broadcast to the observer as a set of visual cues or clues that are intrinsically meaningless. Here the claim is that the relation between clues and assigned meaning is homomorphic: one clue may be assigned many meanings, many clues may be assigned the same meaning. The perceiver, therefore, is assumed to possess a memory-based code that rationalizes the complex of clues read off the retinal image. In sum, where the stimuli are considered as only clues to environmental facts, there is a need for positing mediating constructive activity as the means by which these facts are determined.

Information Processing: A Methodology for Constructivism

Closely cognate with the constructivist philosophy is an approach to problems of perception that is currently in vogue and that may be loosely referred to as "information processing" (Haber, 1969). Implicitly, it takes as its departure point the assumptions as noted above. More explicitly it defines perception not as immediate but as a hierarchically organized temporal sequence of events involving stages of storage and transformation. Transformations occur at points in the information flow where storage capacity constraints demand a recoding of the information. Such recoding must exploit long-term memory structures--or internal models of the world--and, in keeping with a fundamental constructivist belief, perceiving cannot be divorced from memorial processes. Guided by these assumptions, information processing seeks methods that will differentiate the flow of visual information on the nervous system; that is, methods that will decompose the information flow into discrete and temporally ordered stages. In the main, backward masking (e.g., Sperling, 1963; Turvey, 1973), delayed partial-sampling (Sperling, 1960), and reaction-time procedures (e.g., Posner and Mitchell, 1967; Sternberg, 1969) singly or in combination have provided the requisite tools.

An example will provide an elementary illustration of the information-processing approach: the simple task introduced by Sternberg (1966, 1969). A display of one to four characters is presented briefly to an observer who must press a key to indicate whether a subsequently presented single character was or was not a member of the previously displayed set. In general, as the number of items in the memory set increases from one to four, the latency of the observer's response to the probe increases linearly. The linear plot, of latency against number of items, has two characteristics: slope and intercept. We might assume that the slope of the function identifies the process of comparing the probe character with the representations of the characters in memory. But what of the intercept? Does it signify a perceptual/memorial operation or simply the time taken to organize the response?

Suppose that we now degrade the probe character in some way. Does the degrading affect the memory comparison or some other process? If it affects memory comparison, then we should expect the slope to alter, assuming the validity of our original interpretation. An experiment of this kind reveals that degrading the probe essentially leaves the slope of the function invariant but it does raise the intercept (Sternberg, 1967). We can now argue that the intercept reflects, at least in part, that processes of normalizing and perceiving the probe precede memory comparison. In short, in the performance of this simple task we can identify two independently manipulable and successive stages Witness to the potential usefulness of this distinction is the observation that with words as both memory items and probes, poor readers differ from good readers only in the height of the intercept (Katz and Wicklund, 1972).

Though it is true that information processing as an approach often provides an elegant framework and set of procedures for examining perceptual processes, it is also the case that the descriptions it yields are for the most part crude and approximate. It is not unjust to say that the information-processing methodology is limited to a broad identification of stages; and is inherently insufficiently powerful to supply sophisticated descriptions of perceptual procedures and their complex interrelationships. For the achievement of a more rigorous account of the how of perception, constructivism may have to look elsewhere.

Scene Analysis by Machine: Formalized Constructivism

It is by now evident to the reader that constructivism conceives perception as an act involving a potentially large variety of knowledge structures. To gain a purchase on the form of such structures, to discover effective representations (Clowes, 1971) and how they could relate, is in part the task of research and theory in artificial intelligence. It will prove instructive for our purposes to look at systems sufficiently intelligent to infer the three-dimensional structure of objects from two-dimensional line portrayals of opaque polyhedra of the type depicted in Figure 2.

The early work in pattern recognition by machine was dominated by models that held that patterns could be classified by a procedure listing feature values and then mapping these values onto categories through statistical decision processes (e.g., Selfridge and Neisser, 1960). Contemporary work follows the principle that pattern classification systems must possess the ability to articulate patterns into fragments and to specify relations among the articulated fragments (Minsky, 1963). Consequently, artificial intelligence research has been attracted to the structural models that have proved successful in linguistics and the focus of the enterprise has switched from the problem of pattern recognition to the problem of pattern description. The search is for structural grammars that describe the relationships among the parts of a pattern.

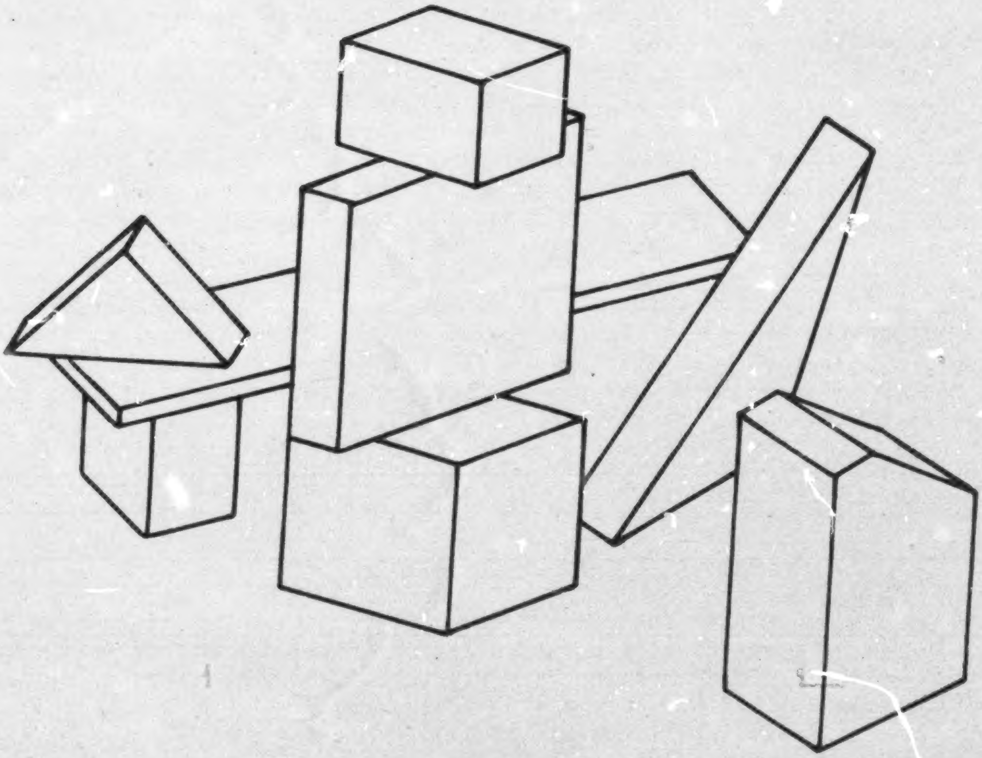

Figure 2: A scene consisting of polyhedral objects. (After Guzman, 1969.)

As a preliminary to our discussion, let us conjecture on the types of stages that might intervene or mediate between a picture and the resulting three-dimensional description. First, we can assume that a picture is projected onto the retina as an array of points that can be partitioned into sets as a function of brightness. The three-dimensional properties of a picture cannot be inferred from this initial points representation. The endeavors, then, of early visual operations must be that of recovering from these patches of brightness, the lines that make up the picture and from these lines, the regions into which the picture may be parsed. (For the purposes of computation, a region may be defined as a set of points with the property that a path drawn between any two elements of the set does not cross a line.) We can usefully refer to these three representations--points, lines, and regions--as being in the picture domain, for as yet they are indifferent to the three-dimensional structure the picture represents. But what we want is a three-dimensional account, a description of a scene. Having recovered regions, the next constructive operation is to map the regions description onto a surfaces description and this in turn onto a description of the bodies to which the surfaces belong. To facilitate the discussion that follows, we will refer to these two representations--surfaces and bodies--as being in the scene domain and recognize that most of artificial intelligence research on intelligent picture processing has been concerned with the mapping between the picture and scene domains. The final representation and domain--the objects domain--is obtained by identifying the objects the bodies represent. This task probably requires knowledge of a higher order such as permissible real-world relations among surfaces and the functional capabilities of variously described bodies (for example, cups are for containing liquids but they could also be used as effective paper weights).

Thus, arriving at a three-dimensional description of a picture may be interpreted as the construction of a number of representations in an orderly fashion from less to more abstract. A representation is said to consist of a set of entities together with a specification of the properties of those entities and the relationships existing among them (Sutherland, 1973). We can now provide an elementary description of computer methods for analyzing complex configurations of objects presented pictorially. The computer is programmed to begin by detecting local entities or features of a given picture. Then it searches for relationships among entities that indicate the presence of particular subpatterns. The determination of specific relationships of subpatterns allows for the detection of more global patterns, i.e., for the establishment of a higher-order representation, and so the procedure continues until a satisfactory structural description of the scene has been obtained.

Let us examine in an approximate way a few representative programs for scene analysis. We are indebted to Sutherland's (1973) lucid clarification of these complex programs.

We begin with Guzman's (1969) well-known program partly because it is relatively simple and partly because it provides the jumping-off point for many subsequent programs. Guzman's program takes an input that has already been parsed into regions and seeks to discover from this description how many separate bodies are present in the two-dimensional portrayal. The goal of the program is quite modest: it aims only at specifying which groups of regions are the faces of a single body--a preliminary step to arriving at a three-dimensional account. The inferences are based on the properties and implications of vertices of the kind shown in Figure 3, with each vertex being classified on the basis of how many

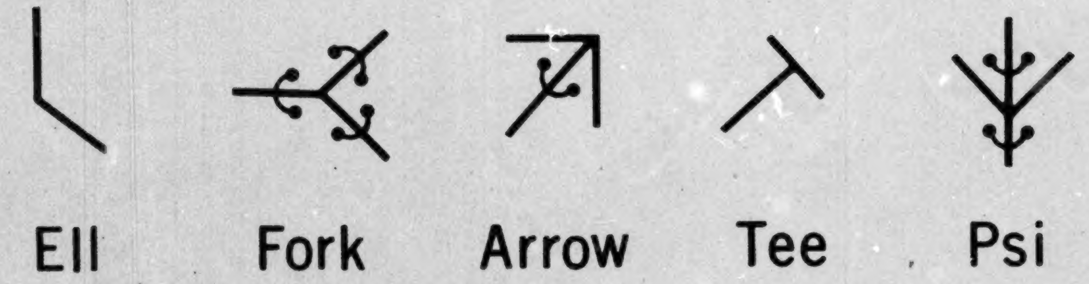
Ell
Fork
Arrow
Tee
Psi

Figure 3: Examples of vertices and the links they imply.

FIGURE 3

lines meet at the vertex and their respective orientations. Essentially, the program having classified the vertices into types, establishes links between regions that meet at vertices. Consider the arrow type of vertex. This would normally be caused by an exterior corner of an object where two of its plane surfaces form an edge. Consequently, the two regions that meet on the shaft of an arrow--that is, those that are bounded by the two smaller angles--are linked; those regions that meet these regions at the barbs of the arrow, however, are not. Similarly, a vertex of the fork type depicts three faces of an object so that links may be inserted across each line. By linking the regions as a function of where they meet, it is possible to separate the bodies represented in Figure 2. The final stage of the program consists of grouping together regions connected directly by a link or indirectly by a chain of links with other regions.

Where a region is connected by only a single link to the members of a connected collection of regions, it is not identified as a member of that collection. Additionally, the program incorporates a policy of examining in a limited fashion neighboring vertices in order to determine whether links should indeed be made at a given vertex.

It is evident that the more closely and formally we examine the question of how a picture is decomposed into separate bodies, the more sensitive we become to the complexity of the problem. Guzman's (1969) program, though primitive, hints at the type of heuristics that may have to be employed by a human observer in arriving at a description of a picture. Of special importance is the suggestion that the figure-ground or segmentation problem, which is often discussed airily in a single sentence of a text on visual perception, is not likely to be solved by the nervous system in any simplistic manner. In this context one is reminded of Hebb's (1949) distinction between primitive and nonsensory unity. Segmentation in the former case could be done on the basis of brightness differentials, in the latter it could only be achieved, so Hebb argues, through an application of knowledge of patterns and objects.

Guzman's program is not especially successful. In a significant number of instances it fails to achieve a correct decomposition of a scene. It is obvious the program is insufficiently knowledgeable. What is needed is knowledge of what is permissible and what is not in the structuring of real-world objects.

To map from regions to surfaces asks that we describe lines in the picture domain as entities in the scene domain. In the general case, a picture line can represent a number of different scene entities, such as edge or boundary of a shadow. How may object edges be recovered from line drawings? It turns out that vertices provide a useful basis for the edge interpretation of picture lines. Regions that meet at a line will correspond to surfaces in the scene domain that meet at a convex or concave edge; certain arrangements of adjacent regions will correspond to the occlusion of one surface by another. Fortunately, with respect to edges each vertex type identifies a limited set of three-dimensional interpretations. Furthermore, by mapping vertices in the picture domain onto corners in the scene domain it becomes possible to eliminate ambiguous edge interpretations: the criterion is whether the set of edge interpretations for a set of vertices around a region is consistent (Clowes, 1971).

Limiting his problem to that of pictures with no more than three lines meeting at a point, Clowes (1971) draws up an exhaustive classification of corners into four types. The classification is based on the concave/convex relationship between surfaces and the visible/invisible properties of surfaces. A Type I corner is defined as one in which the three edges meeting at the corner are all convex; a Type II is one in which two of the edges are convex and one concave; a Type III corner has two concave edges and one convex edge; and a Type IV corner is one in which three concave edges meet. The corners are further subcategorized according to the number of visible surfaces.

How may we construct a representation of a picture in terms of surfaces (three-dimensional entities) and their relationships? In part the solution lies in defining the mapping of vertex type onto visible corner type. First, the Clowes program recovers regions and,associated with each region, the lines and vertices bounding the region. Then the vertices are classified and for each vertex recovered the program lists the possible corner interpretations. The problem now is to arrive at an unequivocal description of corners and hence of the manner in which the surfaces present in the picture articulate. To do this the program asks: Which set of corner interpretations are compatible with the structure of three-dimensional bodies? Obviously, the program must instantiate knowledge about the physical structure of polyhedra in order that it might answer this question and reach a satisfactory three-dimensional description of the scene. It turns out that the following single fact proves sufficient to eliminate impossible combinations of corner interpretations: where two surfaces meet at an edge, that edge must be invariant (either convex or concave, or implicate one surface behind another) throughout its length. Thus, given two vertices connected by a line, one could not interpret one vertex as being of Type I and then interpret the other as being of Type IV, for this would mean that the edge corresponding to the connecting line changes from convex to concave, which is contrary to the structure of possible polyhedra. It is important to recognize that where this injunction imposes constraints on the interpretation of a pair of connected vertices it automatically restricts the corner interpretations of neighboring vertices. Through iteration it will lead to the discovery of sets of compatible corner interpretations.

Thus, the Clowes program exceeds the proficiency of Guzman's by constructing a three-dimensional description of the bodies present in a scene. It successfully differentiates holes from bodies and rejects impossible polyhedral structures. Its limitations include that it accepts only pictures with three or fewer lines converging on a point, and it accepts these pictures only when the lines have been specified beforehand. We may say of Clowes's program that it starts upstream avoiding the problem of recovering lines from points.

Although this may not appear to be overly serious, consider that in naturally illuminated scenes it is often the case that faces of different bodies may reflect the light equally, which means that some edges will not yield a brightness differential or,if they do,it is below some working threshold. How might such edges be recovered? How are lines missing from the picture domain detected?

Where programs have been written to take as their starting point the actual output from a television camera (i.e., an input of points), it has proved fruitful, even necessary, to introduce greater flexibility in the relations among representations and to extend the knowledge of the program to include descriptions (prototypes) of possible objects. This higher-order knowledge can be used to guide the construction of lower-level representations.

A program that makes modest use of these principles is that of Falk (1972). Here directly recoverable lines are assigned to different bodies by a variant of Guzman's program. Special heuristics beyond those specified by Guzman are needed, however, to achieve a segmentation into bodies given that certain lines are just not presented in the input. At all events, having obtained a rough decomposition of the scene into bodies, the program then applies heuristics to fill in the missing lines. It is important to recognize that these low-level operations are meant only to provide a sketch--a working hypothesis--of what a representation of the bodies in the scene might look like.

But in order that we might better understand Falk's program, let us examine more closely the kinds of scenes the program is dealing with and some of the program's esoteric capabilities. There are nine permissible objects (polyhedra, as before) of fixed size that can be variously arranged on a table top such that some objects may be resting on the faces of others. The program possesses highly specialized knowledge about the three-dimensional coordinates of the television camera relative to the table surface, and about the exact size and shape of each of the nine objects. The program will use its exact knowledge of the dimensions of the permissible objects in order to recognize them, but to apply this knowledge, it needs to be able to compute the lengths of at least some of the edges of a given body. Since it knows the position of the camera relative to the table top, it can infer the exact position of any point on the table in 3-space, and in consequence, the lengths of those edges that contact the table can be determined.

The outcome is a determination of which base edges of which bodies contact the table and the program then calculates the lengths of the base edges of each body and the angles between them. These data are then subjected to operations that seek to match the bodies corresponding to the calculated base lengths and angles to the program's knowledge of permissible objects. Recognizing a body as a particular object carries the bonus of knowing fully the exact dimensions of the body--and that includes, of course, the faces that might be "invisible" in the picture. By recognizing which objects are in contact with the table, it is now possible to reassess the earlier conclusions about which bodies were supporting which bodies.

In the final stages of Falk's (1972) program, once supported bodies are isolated from those resting on the table; it maps the supported bodies into the object domain and is then in a position to provide a description of the entire scene in terms of objects and their relationships. But recall that the lines representation from which all else developed was incomplete and the bodies representation was constructed with a strong element of guesswork. To ensure that the description in the objects domain is accurate, a lines representation is synthesized from the objects representation and mapped onto the original lines representation. Where a significant mismatch occurs, the bodies and the objects representations may be revised guided by the now available knowledge about the global properties of the scene.

The latter provides the significant feature of the Falk program, namely, the introduction of a more flexible processing routine, which permits less abstract descriptions to be reexamined and the information present there to be reinterpreted in the light of more abstract descriptions--thus the objects representation is used to reevaluate the lines representation. This mode of organization is becoming more of a necessity in writing computer programs to perform successful scene analysis (Minsky and Papert, 1972). We have seen that the program relies heavily both on exact knowledge of a limited number of objects and on the fact that no other objects than these will be pictured. But we should not be too seriously put off by this austerity and rather should view Falk's program as an illustration of how prototypes might be exploited in visual perception. This approach is represented more elaborately in the program of Roberts (1965), which has knowledge both of prototypes and of their lawful perspectives. Roberts argues that the computer program should assume, like the human observer, that a picture is a perspective of a scene obeying the laws of projective geometry. In addition, the Roberts program uses only three prototypes--a cube, a wedge, and a hexagonal prism--and is able to treat a large array of complex polyhedral objects as combinations of these prototypes. The idea that the structure of human memory for things seen may be described as prototypes together with transformation rules has received some measure of experimental support (Franks and Bransford, 1971).

If I have dealt at some length with these examples of seeing machines, it has been for several reasons. In the first place, they foster an appreciation of the complexity of formalizing visual processes within the constructivist framework and yet at the same time suggest the directions this formalizing might take. In the second place, they rarely find their way into the psychological literature on visual perception and that, in my opinion, is a serious oversight from the point of view of constructive theory. Lastly, they permit the illustration of certain principles that may have significant implications for the general theory of vision.

Thus from the perspective of a seeing machine, the stuff with which a brain works is descriptions, and not images, of optical events. Efforts to explain how vision works ought to focus on computational or symbolic mechanisms rather than the physical mechanisms that have been the mainstay of traditional theories. The latter, as Minsky and Papert (1972) point out, are inherently incapable of accounting for the influence of other knowledge and ideas on perception.

The particular form of computational or symbolic mechanism currently being advanced is not hierarchical. It cannot be said to consist of parts, subparts, and sub-subparts that stand in fixed relation to each other. While a hierarchical label may be appropriate for a physical system, it is less so for a computational system. The latter often exploits the method of two different procedures using each other as subprocedures, and thus what is "higher" at one time is "lower" at another. To capture the essence of nonhierarchical, highly flexible mechanisms, the terms "heterarchical" (McCulloch, 1945; Minsky and Papert, 1972) and "coalitional" (Shaw, 1971; Shaw and McIntyre, 1974) have been suggested.

Though the formalization of coalitions is still very much in its infancy, we can identify in a rough and approximate way some fundamental features that distinguish the coalition (heterarchy) from the more familiar hierarchy. First, many structures would function cooperatively in the determining of perception, although not all structures need participate in all determinations. Second, while it is certainly the case that a coalitional system has very definite and nonarbitrary structures, the partitioning of these structures into agents and instruments and the specifications of relations among them is arbitrary. In short, any inventory of basic constituent elements and relations is equivocal.

Perhaps the main emphasis of the coalitional formulation is the flexibility of relations among structures. Falk's (1972) program is a very modest instantiation of these coalitional features, Winograd's (1972) celebrated language understanding system is a more ambitious one. Contrary to much of current theoretical linguistics, Winograd's system is concerned more with the problems of representing the meanings conveyed by discourse than with the grammatical structure of discourse. The system is predicated on the coalitional thesis that sentence comprehension necessitates an intimate and flexible confluence among grammar, semantics, and reasoning. In the Winograd system the sentence "parser" can search out semantic programs to determine if a particular phrase makes sense; semantic programs can exploit deductive programs to determine whether the proposed phrase is sensible in the context of the current state of the real world (Minsky and Papert, 1972; Winograd, 1972). The fundamental principle of operation, though complex, may be stated simply: each piece of knowledge can be a procedure and thus it can call on any other piece of knowledge.

Perception at a Glance: The Contribution of the Information-Processing Approach

The preceding discussion has focused on the theory of seeing as a constructive act. In an elementary but sufficient manner, we have considered some of the principal notions, the scaffolding if you like, on which theories of indirect perception are erected. Next let us examine experimental efforts to unravel the processes by which information to an eye is "transformed, reduced, elaborated, stored, recovered and used" (Neisser, 1967:4).

Our initial focus is the analysis of perception at a glance. Much of information-processing research has revolved around tachistoscopic presentations and an interpretation of their perceptual consequences. Moreover, the materials briefly exposed have been for the most part letters, numbers, and words, so that we might take the liberty of describing the analysis as that of "the stages underlying the perception of linguistic material in a single fixation." The dominant use of linguistic material inhibits the elaboration of tachistoscopic perception into a general theory of visual-perception-at-a-glance, but given the language-processing interests of this conference that limitation is perhaps of no great consequence. Let us remind ourselves that on the constructivist view an understanding of perception in a single fixation is fundamental since normal everyday perception is construed as the fitting together of successive retinal snapshots.

Iconic and Short-Term Schematic Representations

There is one thing we can assume from the outset: if perception is a process over time, then in the cases of brief but perceivable optical events (say of the order of several milliseconds, to provide an extreme instance), there ought to be a mechanism that internally preserves such events beyond their physical duration. It is on this internally persisting representation that constructive operations are performed and in its absence we ought to suppose that the perception of brief displays would be well nigh impossible. Furthermore, we can suggest the following about this representation: it should be in an uncategorized form. Suppose that the mechanisms of perception were prefabricated in the sense of possessing knowledge about, and routines for, abstracting universal regularities, and that these regularities were automatically registered and classified. One consequence of this arrangement would be category blindness (MacKay, 1967). If a new category attained significance for the organism, the organism would not be able to grasp it. Obviously, the perceptual mechanisms, although partly prefabricated, must also be flexible. What we can imagine is that on the occurrence of an optical event prefabricated general procedures are brought to bear, and a variety of testings are carried out to determine how this event might fit into the current organization. This suggests (see MacKay, 1967) that there ought to be a region or "workshop" in the perceptual system that allows for hypothesis testing before categorization. We might anticipate, from the constructivist position, the existence of a transient memory for visual stimulation that preserves the raw data in a relatively literal form.

In many respects the pivotal concept in the information-processing account of perception-at-a-glance is a memory system having the properties suggested above. The original evidence for a transient, high-capacity, literal visual memory comes from the well-known experiments of Sperling (1960) and Averbach and Coriell (1961).

Brief visual storage, or iconic memory as Neisser (1967) has termed it, was isolated through the use of a delayed partial-sampling procedure. Essentially, this procedure involves presenting simultaneously an overload of items tachistoscopically followed by an indicator designating which element or subset of elements the subject has to report. If the indicator is given soon after the display, the subject can report proportionately more with partial report than if asked for a report of the whole display. This superiority permits the inference of a large capacity store; the sharp decline in partial-report superiority with indicator delay permits the inference of rapid decay. Purest estimates of the decay rate reveal that the duration of this storage is of the order of 250 msec (Averbach and Coriell, 1961; Vanthoor and Eijkman, 1973).

The proof of the precategorical character of iconic storage is found in the kinds of selection criteria that yield efficient performance in the delayed partial-sampling task. Generally, superior partial report can be demonstrated when the items in a display are selected for partial report on the basis of brightness (von Wright, 1968), size (von Wright, 1968, 1970), color (von Wright, 1968, 1970; Clark, 1969; Turvey, 1972), shape (Turvey and Kravetz, 1970), movement (Triesman, Russell, and Green, 1974), and location (e.g., Sperling, 1960). Partial-report performance, however, is notably poorer (i.e., not significantly better than whole report) when the letter/digit distinction is the basis for selection (Sperling, 1960; von Wright, 1970). On these data we can conclude that one can select or ignore items in iconic storage on the basis of their general physical characteristics, but one cannot with the same efficiency select or ignore items on the basis of their derived properties. We might wish to argue, therefore, that the iconic representation is literal.

We can gain a richer understanding of the character of iconic storage by comparing it with another form of visual memory that arises quite early in the flow of information and that may be described as abstract or schematic. As we shall see, the longevity of this representation exceeds by a considerable degree that of the icon.

It has often been argued that the iconic representation of linguistic material undergoes a metamorphosis from a visual to a linguistically related form. One elegant expression of this view suggests that the raw visual data are cast rapidly into a set of instructions for the speech articulators for subsequent (and more leisurely) rehearsal and report (Sperling, 1967). However, as one would intuit, the transformation of the icon establishes not only a representation in the language system but, in addition, and we may suppose in parallel (Coltheart, 1972), a further and more stable representation in the visual system. The first strong experimental hints that this might be so were provided by Posner and Keele (1967) in an experiment that stood on the shoulders of an earlier series of now celebrated experiments by Posner and Mitchell (1967).

Consider a situation in which subjects are presented a pair of letters and asked to respond "same," if these letters have the same name, and "otherwise," if they are different. On some occasions the letters with the same name are also physically the same (e.g., AA) and on others the letters with the same name are physically dissimilar (e.g., Aa). It proves to be the csse that "same" responses to AA are significantly faster than same responses to Aa, which suggests that AA-type pairs are not necessarily being processed on the basis of name but rather that their visual characteristics are being used to make the response. Adopting Posner's (1969) terminology, we will refer to matches of physically identical letters as "physical matches" and matches of physically different but nominally identical letters as "name matches."

We now look at what happens when the two members of a pair are presented sequentially rather than simultaneously and the time elapsing between the appearance of the first and the appearance of the second is varied. Under these conditions, the latency of a name match is indifferent to the delay time, but the latency of a physical match increases with delay until it and the name match latency are virtually identical. The converging of physical- and name-match latencies identifies a decline in the availability of a visual code and an increasing dependence on the name of the letter with the passage of time. In the original experiment (Posner and Keele, 1967), the estimated duration of the visual code isolated by the reaction-time procedure was of the order of 2 sec; however, subsequent research has shown that it is considerably more durable than the original experiment would have us believe (e.g., Kroll, Parks, Parkinson, Bieber, and Johnson, 1970; Phillips and Baddeley, 1971). At all events, the superior longevity of this visual code to that exposed by the delayed partial-sampling procedure (namely, the icon) suggests that we are indeed dealing with two different visual, memorial representations--two different descriptions--of an optical event. But before elaborating on this conclusion, we ought to be more explicit about the difference between the two procedures defining the two representations. In one, delayed partial sampling, we are interested in the persistence of aspects of visual stimulation not yet selectively attended to; in the other, our concern is with the persistence of the visual description of stimulation that has enjoyed the privileges of selective attention.

Beyond the difference in persistence between the two visual representations we can note the following differences. In the first place, the iconic representation is perturbed by an aftercoming visual mask, the schematic representation is not (cf. Sperling, 1963; Posner, 1969; Scharf and Lefton, 1970; Phillips, 1974). In the second place, the capacity of the iconic representation is probably unlimited, but that of the schematic representation is clearly constrained (Coltheart, 1972; Phillips, 1974). In the third place, if during the existence of a representation demands are made concurrently on processing capacity (Posner, 1966; Moray, 1967), the persistence of the iconic representation seems to be unimpaired (Doost and Turvey, 1971) in contrast to the persistence of the schematic representation, which can be shown to be severely reduced (Posner, 1969). The corollary to this latter distinction, however, is that the close dependency of the schematic representation on central processing capacity means that the persistence of this representation is indeterminate--it may persist for as long as sufficient processing capacity is devoted to it (Posner, 1969; Posner, Boies, Eichelman, and Taylor, 1969; Kroll et al., 1970).

One last difference is worth attention. The high-capacity, maskable iconic representation is tied to spatial position; the low-capacity, unmaskable schematic representation is not (Phillips, 1974). On the basis of the schematic representation, the perceptual system can match two successive optical events when they are spatially separate as efficiently as it can when they are spatially congruent. On the basis of the iconic representation, however, a match of successive and spatially separate events is conducted less efficiently than a match of successive events that spatially overlap (Phillips, 1974).

We could go farther in our discussion of these two largely different descriptions of the visual appearance of a stimulus but what has been said is sufficient for our purposes. Let us therefore conclude this discussion with a crude sketch of the role played by the iconic description in the general scheme of visual information processing.

For current theory, memory consists of two basic structures, termed "active" and "passive," or alternatively two basic models, the "short-term" and the "long-term." The relation of the iconic representation to these two structures is not perfectly obvious, although an appeal to the general consensus of opinion (Neisser, 1967; Sperling, 1967; Haber, 1969; Turvey, 1973) informs us that the icon is a necessary precursor to active memory and that it interfaces visual stimulation with internal models. Thus in this view the iconic representation is a transient state that, in the hands of long-term knowledge structures, is molded into a variety of representations in active memory. On the foregoing account, we may identify the schematic visual code and the name code as examples of active representations so produced. Figure 4 captures these ideas.

Temporal Characteristics of the Iconic Interface

How might we examine the fine temporal grain of events surrounding the iconic interface? One tool that has proved reasonably successful in this regard is visual masking--the impairment in the perception of one of a pair of stimuli when the two are presented in close temporal succession. There are a number of renderings of this phenomenon, some of which are particularly significant to the discussion that follows. First, the influence of the second member of the pair on the first, which is referred to as "backward masking"; and the influence of the first, on the second, which is termed "forward masking." Second, the two stimuli may be presented binocularly, i.e., both stimuli are presented to both eyes, or monoptically, i.e., both stimuli are presented to one eye; or may be presented dichoptically, in which case one member of the stimulus pair is presented to one eye and the other member is presented to the other eye. Masking that occurs under conditions of binocular or monoptic viewing may originate in either peripheral or central visual mechanisms, but masking that occurs in

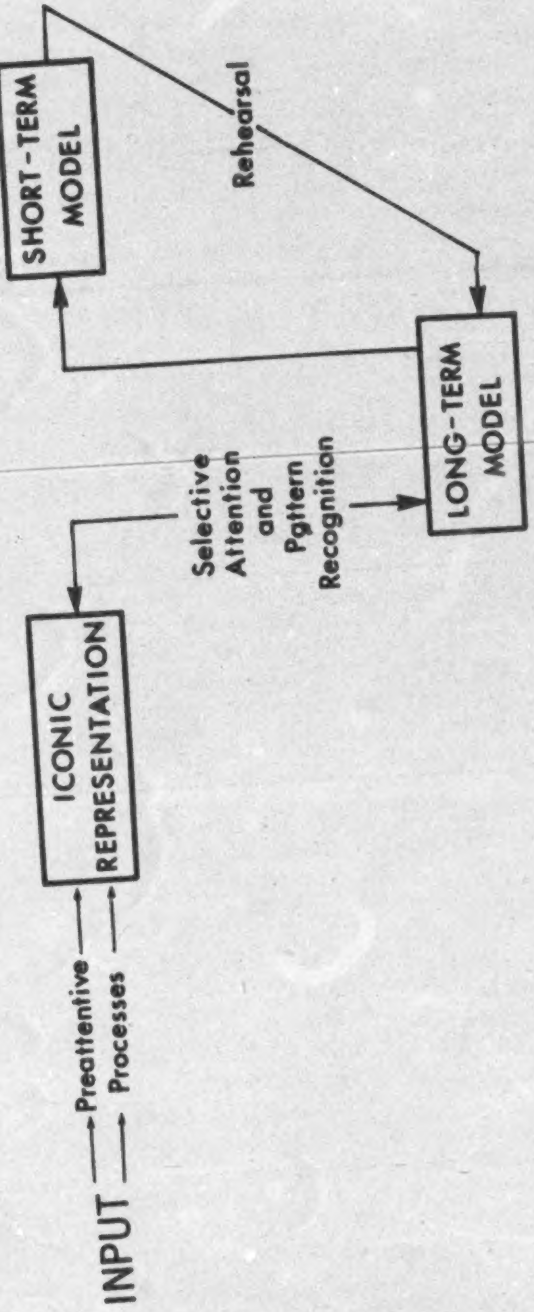
INPUT
Preattentive Processes
ICONIC REPRESENTATION
Selective Attention and Pattern Recognition
LONG-TERM MODEL
SHORT-TERM MODEL
Rehearsal

Figure 4: A typical information-processing scheme.

FIGURE 4

dichoptic viewing is more likely due to central effects. Third, the two stimuli may or may not overlap spatially. We reserve the special term "metacontrast" for the case in which backward masking occurs with nonoverlapping but spatially adjacent stimuli.

Essentially, there are two major interpresentations of masking effects. If we refer to the stimulus to be identified as the target and the stimulus impeding perception as the mask, then one interpretation, the integration hypothesis (Kahneman, 1968), stresses the effect that the mask has on the target representation. The idea is that the two stimuli that follow one another in rapid succession are effectively simultaneous within a single "frame" of psychological time, analogous to a double exposure of a photographic plate. The result of this summation favors the higher-energy stimulus; thus, if the mask is of greater energy, then the target will be reduced in clarity and its identification impeded. Closely related to this interpretation is the notion that where a greater-energy mask follows a target the neural response to the target is occluded. This view is often dubbed "overtake" and it may be thought of as an integration emphasizing a nonlinear summation of responses rather than a linear summation of stimuli (Kahneman, 1968).

A contrasting interpretation of masking is presented in the form of the interruption hypothesis (Kahneman, 1968): if a mask follows a target stimulus after some delay, processing is assumed to have occurred during that delay but is terminated or interfered with by the mask. In the context of the iconic representation, we can view the interruption hypothesis as saying that an aftercoming stimulus does not affect the accuracy of the iconic description of a prior stimulus but rather interferes with its translation into active memory codes. In this same context it is evident that the stimulus version of the integration hypothesis indicates that the target and mask are dealt with as a composite, resulting in an iconic representation in which the target is unintelligible. The response version of the same hypothesis suggests that the iconic description of the target is never formed.

When looked at in terms of the icon, the two interpretations of masking do not appear as competing explanations, rather they seem to be interpretations of masking originating at different stages in the flow of visual information (Turvey, 1973).

We can provide some measure of proof for this point of view through an experiment of the following kind. A target stimulus, say a letter, is presented briefly (<10 msec) to one eye followed shortly afterward (say, 0 to 50 msec) by a similarly brief exposure of a masking stimulus to the other eye. A contoured mask with features in common with the target will seriously impair the perception of the target in this dichoptic situation but, by way of contrast, a noncontoured mask that bears no formal relation to the target, such as a homogeneous flash or a fine-grain random dot pattern, will not; even if within limits, its energy is made considerably greater than that of the target (Schiller, 1965; Turvey, 1973). Yet if a noncontoured mask is presented to the same eye as the target, masking can be demonstrated. On these observations it may be aruged that with respect to a contoured target, a contoured mask can exert a central influence but a noncontoured mask cannot. The influence of the latter is limited primarily to the peripheral processing of the target.

The gist of the whole matter is given in an experimental arrangement in which a contoured target and an aftercoming contoured mask are presented to separate eyes with a noncontoured higher-energy mask following shortly thereafter on the same eye as the contoured mask. The perceptual outcome of this apparently complex configuration of events is singularly straightforward: the target can be seen and identified against the unimpeding background of the noncontoured mask (Turvey, 1973). Although straightforward, the outcome is curious. It implies that the second arriving mask occluded the first arriving mask in the "peripheral" sequence of neural transformations from retina to cortex. That is to say, the first mask that could impede target perception centrally never in fact reached central processing mechanisms and the perceptibility of the target was thus unhindered. Thus, one might venture to propose that different kinds of masking obeying different principles originate at different stages in the flow of visual information on the nervous system

Consider the monoptic masking of letter targets by a masker of equal or greater energy that is relatively ineffectual dichoptically. It has been demonstrated that in this situation the following principle relates target duration to the minimal interstimulus (target offset to mask onset) interval permitting evasion of the masking action: target duration × minimal interstimulus interval = a constant (Turvey, 1973; Novik, 1974). We owe the first demonstration of this relationship to Kinsbourne and Warrington (1962a).

When examined more closely, target energy rather than target duration emerges as the true entry into the relation (Turvey, 1973). Furthermore, it is of some importance that the relation holds for forward as well as backward masking, although the constant for the forward relation is higher than that for the backward (Kinsbourne and Warrington, 1962b; Turvey, 1973); in brief, forward masking in the domain of the multiplicative rule extends over a greater range than backward masking.

My own treatment of the multiplicative function is that it characterizes peripheral processing. If we assume that there are a large number of independent and parallel networks detecting features and/or spatial frequencies (see below), then the relation tells us, or so I argue, that the rate of detection is a direct function of target energy.

We should inquire as to the relation between target duration and the minimal interstimulus interval when the two stimuli are presented dichoptically and the masker is a contoured, effective dichoptic-masking agent. In this situation, the relation proves to be additive: target duration + minimal interstimulus interval = a constant (Turvey, 1973; Novik, 1974) and here target duration is the proper entry. Indeed, this dichoptic or central principle tells us that the total time elapsing between stimulus onsets is what is significant. While the energy relation between target and mask is of major importance to peripheral processes, it is of limited importance to central processes.

Perhaps the brunt of the peripheral/central difference is carried by the contrast of forward and backward masking. We recall that peripherally forward masking was greater than backward even though both obeyed the same principle. Contrapuntally, central forward masking is slight in comparison to central backward masking (Smith and Schiller, 1966; Turvey, 1973). We see that central masking is primarily backward and this may be an important comment on the nature of central information processing (cf. Kolers, 1968).

When two successive stimuli compete for the services of the same central processes, it is the later arriving one that is more completely identified. On the other hand, when two stimuli compete for the same peripheral networks, order of arrival is less important than energy. Peripherally the stimulus of greater energy, whether it leads or lags, will be the one whose properties are likely to be registered.

We ought to ask how the two rules described above relate to each other. On the hypothesis that the multiplicative rule relates to primarily peripheral processes and the additive to central, our question becomes that of how peripheral and central processes combine. At first blush we might think that this relationship is successive and additive; that is to say, peripheral processing is completed first, then central processes occur, and total processing time is given by adding the two stages. It proves to be the case, however, that a successive and additive interpretation will not do (Turvey, 1973: Exp. 12). A more reasonable interpretation is that the output from peripheral networks is parallel and asynchronous--different features are detected at different rates (Kolers, 1967; Weisstein, 1971)--and that central processes, obviously dependent on peripheral input, operate simultaneously with peripheral processes. The relation is said to be concurrent and contingent (cf. Turvey, 1973).

Identifying Feature-Detectors in Human Information Processing

We noted earlier that "information processing" as a methodology for examining transformations of neural states may not be able to give us a measure of vision's structural grammars nor a sufficiently detailed account of how representations commute. We may now turn to a particularly significant contribution of the approach: its ability to suggest or demonstrate feature-sensitive systems in human vision.

It is well-known that there are cells in the visual cortex of the cat and monkey selectively sensitive to the orientation of lines. How might we reveal the presence of such units in human vision? A demonstration is important because our constructive theories of human pattern and object recognition presuppose the existence of these units.

A phenomenon well-suited to this purpose is the "aftereffect." The logic behind visual aftereffects as a technique is that prolonged viewing of a stimulus consisting of a particular visual feature should selectively depress the mechanism detecting that feature. We can take advantage of a fairly general rule: if any form of stimulation is continued for a long time and then stopped, one will tend to experience the reverse condition.

Orientation specific systems in human vision are hinted at by the tilt aftereffect (Gibson and Radner, 1937). Suppose that one was exposed for a period of time to a high-contrast grating tilted slightly to the right of vertical. A subsequently exposed vertical grating (or vertical line) would then appear tilted slightly to the left. This is termed the "direct effect." The fact that exposure to a vertical grating tilted clockwise will cause a horizontal grating or line to appear counterclockwise from the horizontal is termed the "indirect effect" (see Coltheart, 1971). An eminent example of this aftereffect is provided by Campbell and Maffie (1971). The evidence is that the human visual system exhibits a narrow orientational tuning of the type reported for units in the cat and monkey (Hubel and Wiesel, 1968).

A particularly intriguing aspect of the tilt aftereffect is that it can be color-specific. If one is first exposed to red stripes tilted clockwise off vertical and green stripes tilted off vertical by the same amount but counterclockwise and then one examines a vertical test stripe, its apparent orientation will depend on the color of the light in which it is projected. Projected in red light, it will appear tilted counterclockwise; projected in green light, it will appear tilted clockwise (Held and Shattuck, 1971). This orientation aftereffect specific to color should be contrasted with the now famous McCullough phenomenon, which is a color aftereffect contingent on orientation (McCullough, 1965). In the latter, viewing alternatively a vertical grating on a blue background and a horizontal grating on an orange background induces an orange afterimage when vertical lines on a white ground are inspected and a blue afterimage when horizontal lines on a white ground are inspected.

A variety of contingent aftereffects has now been demonstrated. A representative but not exhaustive list would look like this: motion-contingent color aftereffects (Stromeyer and Mansfield, 1970); color-contingent motion aftereffects (Favreau, Emerson, and Corballis, 1972); texture-contingent visual motion aftereffects (Mayhew and Anstis, 1972; Walker, 1972) and curvature-contingent color aftereffects (Riggs, 1974).

Let us consider one further example of how the aftereffect phenomenon demonstrates that different characteristics of visual stimulation can be processed selectively. If one views a contracting simple arithmetic ("Archimedes") spiral for some period and then directs the eyes to another object, one experiences an equally great but opposite effect, i.e., the object will appear to expand. But suppose that the adapting spiral subtended only 4° of visual angle and that following prolonged viewing one looked at a large and regular 20 × 20 matrix of squares. A region of the matrix subtending about 4° will appear to enlarge and even to approach. The lines, however, will not look curved nor will there be any interruption between the growing part of the matrix and the remainder of the matrix. As Kolers (1966, 1968) remarks, a part of the figure seems to change in size and position without looking discontinuous with the remainder.

Although aftereffects are curious and engaging, their usefulness as a source of information about feature analyzers in human vision has not been endorsed uncritically by all students of perception. Weisstein (1969) sees the principal weakness of the technique as that of examining what is left after adaptation. It is reasonable to conjecture, and indeed the contingent aftereffects bear this out, that the aftereffect manifests a pooling of many different types of analyzers. Moreover, a number of different populations of analyzers could give the same phenomenal result. Other students have raised similar doubts about the integrity and fruitfulness of aftereffect data for the exploration of feature detectors in human vision (Harris and Gibson, 1968; Murch, 1972).

An alternative strategy is one that is referred to as cross adaptation. The essential characteristic of cross adaptation is that one examines the loss of sensitivity to one pattern given a preceding exposure to another. The measure of change in sensitivity contrasts cross adaptation with the "aftereffect," which is concerned with the degree to which perception is reversed. When a grating is viewed for some period of time, the thresholds for the same and similar gratings are raised, but thresholds for gratings differing in orientation and size are virtually unchanged (Blakemore and Campbell, 1969). This may be taken as evidence that different populations of neurons respond differentially to features of a stimulus.

There is an especially provocative application of this procedure that allows us to draw together several strands of this discussion. Weisstein and her co-workers (Weisstein, 1970; Weisstein, Montalvo, and Ozog, 1972) were motivated by an aspect of the theory of object recognition that is prominent in scene analysis programs but neglected in psychological experiments. In scene analysis, constructing a representation involves abstracting certain entities, identifying their attributes, and specifying the relationships among them. Virtually all aftereffects and all cross-adaptation experiments are directed toward isolating and defining entities. But of interest to Weisstein (1970) was the possibility of demonstrating relations among entities in the scene domain. She inquired whether one could selectively adapt the neural structures responsible for the relation "in back of." Her experiment was deceptively simple. Subjects inspected a vertical grating that was partially blocked from view by a perspective drawing of a cube. A subsequent vertical test grating was presented within the portion of the visual field where the prior grating was visible and also where it was not (i.e., in the region covered by the cube). A reduction in apparent contrast (adaptation) was found for both positions of the test grating. This means that a population of size-selective and orientation-selective systems that was adapted out by the original grating was also adapted out (although not to the same degree) by the cube. Yet it could be domonstrated that a cube by itself (i.e., not covering a grating) does not induce a significant adaptation effect nor does a hexagon outline drawing of a cube that partially occludes the same area of grating (Weisstein et al., 1972). Apparently, what is important to the effect is the impression of depth given by the perspective drawing of the cube, and one might interpret this result to mean that the relation "in back of" is specified by the firing of those cells that would have fired if the grating had been visible in the region covered by the cube. In Weisstein's view, this effect implies the involvement of neural mechanisms that separate a scene into its components--analogous perhaps to the operations we witnessed in the computer programs described earlier.

Information Processing as a Coalitional Skill

We can conclude this brief and selective account of constructivism with an emphasis on the coalitional/heterarchical conception of how perceptual procedures are organized. What follows is a potpourri of curious experimental observations that implicate (but do not necessarily dictate) the form of organization encountered in our earlier discussion of seeing machines.

Oriented-line detectors play a significant role in theories of object recognition. Generally they are assigned to an early stage in a hierarchically organized processing scheme. However, their actual status within the scheme and the scheme's structure are less than obvious, as witnessed by an experiment of Weisstein and Harris (1974). They demonstrated facilitation of "feature" (line) detection by object context. This work is matched by a facilitation of "object" detection by scene context elegantly revealed in the work of Biederman and his colleagues (Biederman, 1972; Biederman, Glass, and Stacy, 1973; Biederman, Rabinowitz, Glass, and Stacy, 1974). An object is more accurately identified when part of a briefly exposed real-world scene than when it is part of a jumbled version of that scene, exposed equally briefly.

These observations are puzzling on the assumption that the detection of fragments of a pattern predates the determination of the global structure and identity of a pattern. Thus, paradoxically, the identity of the whole depends on the identity of its fragments, but the perceptibility of a fragment is determined by the whole in which it is embedded as an integral part. Significantly, where a fragment is immaterial to the global structure its presence is more likely to be obscured than enhanced (see Rock, Halper, and Clayton, 1972).

This paradox is also apparent in the perception of linguistic material. It can be demonstrated that a letter is perceived more readily in the context of a word than in the context of a meaningless string of letters (e.g., Reicher, 1969; Wheeler, 1970; Johnston and McClelland, 1974). Though some have suggested that results of this kind can be interpreted solely in terms of the superior orthographic/phonologic regularity in the word (cf. Baron and Thurstone, 1973), others have sought to demonstrate the significance of meaning. Controlling for orthographic/phonologic regularity (as best one can), Henderson (1974) has shown that meaningful letter strings, such as VD, LSD, YMCA, are compared faster in a binary classification task than meaningless strings, such as BV, LSF, YPMC.

There is some motivation for interpreting results of the latter kind in terms of a direct accessing of semantic knowledge. For example, it has been shown that the time to reject a meaningful consonant triplet as a word is longer than the rejection latency for a meaningless consonant triplet (Novik, 1974). This observation contradicts the idea that an analysis of orthographic/phonologic regularity precedes entry into the lexicon and suggests to the contrary that lexical access is temporally contiguous with structural analysis. One conception that may be useful here is that of Henderson (1974): feature analysis has parallel access to various memory structures or domains (in the vernacular of our earlier discussion)--grapheme knowledge, orthographic rules, a content-addressable semantic base--though consulted in parallel, the domains interact coalitionally through rules that map from one to the other. Presumably perceptual decisions are reached through this mutual cooperation among separate domains.

It is worth remarking further on the idea of a direct mapping between script and meaning. Several scholars have found this notion necessary to their accounts of skilled reading (e.g., Bower, 1970; Kolers, 1970), and there are a number of provocative clinical observations that speak in its favor. To take but one example, a paralexic error (though not a common one) is to read a word as a semantic relative; thus hen is read as "egg" (Marshall and Newcombe, 1971). The reader in this instance cannot identify the word, nor can he give a phonetic rendering of it, but he can relate to its semantic structure. There are parallels to this clinical observation to be found in the visual masking literature, namely, experiments suggesting that an observer may have some knowledge of the meaning of a masked word even though he may be unable to report the actual identity of the word (Wickens, 1972; Marcel, 1974).

There is a second batch of curious results, much related to those just described. Suppose that one is asked to scan a list of items in search of a specified target item. As a first approximation we can say that the perceiver looks for the set of features (or an appropriate subset) that defines the target item. In this case, nontarget items, or foils that share many of the target's features, will be a greater hindrance to the search than foils that are less closely related. Experimental evidence tends to bear this out (Neisser, 1967). Unfortunately, this first aprroximation is challenged by visual scan experiments in which the target is drawn from a conceptual category different from the foils--the target, say, is a letter and the foils are digits, or vice versa. Here the evidence is that the time to find any letter (digit) in a list of digits (letters) is less than the time it takes to find a particular digit (letter) in a list of digits (letters) (Brand, 1971; Ingling, 1972). The implication is that category discrimination can precede character identification: one can know that a character is a letter or digit before one knows what letter or digit it is. Our puzzle is: What are the features that define the category of letters on the one hand and digits on the other? This puzzle is compounded by the following experiment, which plays on the ambiguity of the character "0" (it may be interpreted as either the digit zero or the letter "oh"). When "0" is embedded in a list of digits, it can be found more rapidly if the observer is told that she is looking for a letter than if she is told that she is looking for a digit. Conversely when "0" is a member of a list of letters, latency of search is considerably shorter if one is looking for the digit zero than if one is looking for the letter "oh" (Jonides and Gleitman, 1972). This result is not restricted to the digit/letter distinction, for research at the University of Belgrade reveals the same pattern of findings when the target is an ambiguous letter, that is, a letter having one phonetic interpretation in the Cyrillic alphabet and another in the Roman alphabet.[1] (The popular use of two alphabets is an interesting feature of the Serbo-Croatian language.)

Though we cannot provide an explanation for the phenomena just described, we can appreciate their implications for the modeling of the human as an information processor. Whatever procedures are presumed to be involved in the processing of information, it can be hypothesized that their manner of interrelating is not obligatory. The nature of the task constrains the structure of the coalition with different tasks requiring different coordinations of procedures. In this sense, information processing is a coalitional skill.

PERCEPTION AS PRIMARY AND DIRECT: THE GIBSONIAN ALTERNATIVE

Introduction

Since constructivism starts from the assumption that stimuli are informationally inadequate, then it is obvious why the primary concern of perceptual theory is taken to be the investigation of the how of perception. Constructivism encourages an inquiry into memory structures and cognitive operations that mediate cues, punctate sensations, features, or whatever, and the perceptual experience. On this view internal models enable adaptation to a world poorly signaled by the flux of energy. Thus it is the internal models, their acquisition, and their usage, that we seek to understand.

In polar opposition to this strategy stands Gibson's suggestion that we ask not what is inside the head--as the constructivists would have it--but rather what the head is inside of (Mace, in press). For Gibson the question of the what of perception has been given short shrift, and in his view we are burdened with excessive theoretical baggage relating to the how of perception for the very reason that what there is to be perceived has not been seriously examined (Shaw and McIntyre, 1974).

---

[1] Georgije Lukatela, 1975: personal communication.

Gibson begins with the question: What do terrestrial environments look like? A departure point curiously ignored by those who would build general theories of visual perception. Indeed, it can be argued that the constructivist approach to vision has not broken kinship with the school of thought that gave rise to Molyneux's premise in the seventeenth century (Pastore, 1971). By taking empty Euclidian space as the frame of reference, the intellectual predecessors of modern constructivism were forced into the position of arguing that distance, for example, could not be apprehended through vision. Distance could only be arrived at (constructed, inferred) through the supplementary information provided by past experience and represented in the form of kinesthetic and tactile images. A moment's examination of the classical exposition of space perception will stand us in good stead.

In the classical view the third dimension was construed as a straight line extending outward from the eye. But since physical space was interpreted as an empty Euclidean space, nothing existed between the eye and an object fixated. One might refer to the theory derived from this conception as the "air theory" of space perception (Gibson, 1950), for it implies an observer looking at unsupported objects hanging in midair. (We can readily admit to the unnaturalness of this characterization, but our criticism must be held in abeyance for the time being; a great deal of modern theory has been erected on this way of describing perceptual situations in the abstract, and it is incumbent upon us to appreciate fully the point of view.)

On the "air theory" formulation, an object fixated by an observer projects a two-dimensional form on the retina that relates to the size and outline of the exposed face(s) of the object. What needs to be explained is how the object's distance is perceived. To grasp the nature of this problem, consider Figure 5, which portrays in traditional fashion a number of objects at varying distances from the observer. The size of each object's projection onto the retinal surface is a function of the visual angle formed by the light rays from the extremities of the object. We have chosen our objects and their slants such that the visual angle projected by each is the same. Clearly, size on the retina does not unequivocally specify distance and we are led to conclude that retinal information, and thus vision, is insufficient for object-distance perception. It necessarily follows that for perceiving the distance of objects other information must be supplied, supposedly from the observer's memory banks. For instance, the transactionalists (e.g., Ittleson, 1960) looked to one's familiarity with objects (and thus with their actual dimensions) as the relevant memory material. But this left unsettled the problem of perceiving the distance of unfamiliar objects. For an alternative one could look to the information available in the converging of the two eyes (cf. Gregory, 1969). Supposedly the angle set between the eyes when converging on an object is a cue to its distance. But there is a good reason to doubt the efficacy of convergence (Ogle, 1962) and in any event such a cue is not at the disposal of many animals, namely, those with nonconverging eyes who do perceive object distance. Furthermore, we can readily imagine the problem pictured in Figure 5 as being that of the limited information to a single eye and we can, after all, perceive distance monocularly.

In his 1950 text Gibson responded to the classical treatment of space perception in a way that was both elegant and simple. To begin with he translated the abstract question, "How is space perceived?" into the biologically and ecologically meaningful question of, "How is the layout of surfaces detected?" More

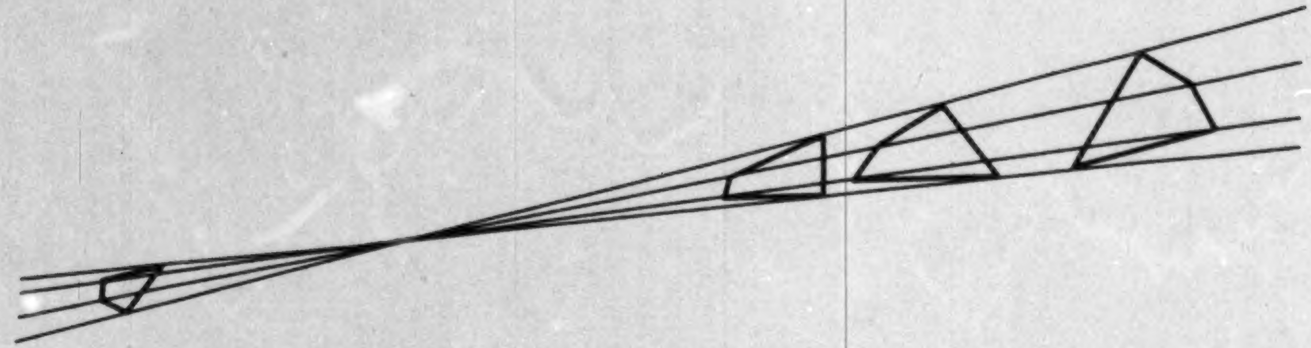

Figure 5: Different forms at different slants and distances from the eye can produce the same retinal image (the form at the left). Thus the retinal image is ambiguous for form, size, and distance.

FIGURE 5

especially, Gibson asked, "How, from a point of observation, do we see continuous distance in all directions?" rather than the commonplace experimental question of "How do we judge the distance of these two objects?" The importance of an impression of continuous distance is that it underlies our capability to perceive the distances of any number of objects in a field of view.

In the next step, Gibson replaced the traditional conception of an isolated eye viewing mathematical points in empty space with that of an eye attached to a body in contact with a ground surface viewing points attached to the surface. In this account of the problem it is evident that by the laws of linear perspective the retinal image is structured in a way that corresponds unambiguously to the distribution of points. (Figure 6 contrasts the classical and Gibsonian accounts.) Moreover, if we now replace the points with objects, then it is similarly evident by the laws of linear perspective that the projection of these objects-on-a-surface will be such that near objects will be imaged large and high up on the retina, while far objects will be imaged smaller and lower down.

Although what has been said so far does not do justice to Gibson's current thinking (particularly the references to retinal image), it is clear that on this reinterpretation of the space perception problem the information to an eye is conceivably richer than traditional theory would have us believe. Indeed, one might venture to say that the light to an eye could be structured in a way that corresponds to the layout of points and objects on a ground surface.

Let us return to the classical view in order to illustrate its pervasive influence on vision theory. Earlier we examined the motivation for assuming that perception goes beyond what is given, that is, the view of perception as indirect. In the classical theory of space perception we have the original impetus: the third dimension is lost in the two-dimensional retina. The immediate theoretical consequence of this "loss" is obvious and we have already deliberated on it. But it is worth remarking once more: the retinal image is held to be a flat patchwork of colors, more precisely of colored forms, to which one can add a third dimension by using available clues, some of which are given directly and some of which are themselves constructions (e.g., superposition).

Of the many conceptual progeny sired by the classical story the following two are among the most significant. First is the notion, now inviolate, that the two-dimensional retinal image is the proper starting point for any theory of seeing. It is this notion that legitimizes, among other things, the enterprise of building theories of visual perception on the shoulders of experiments in picture perception. Moreover, since perceiving begins with a flat patchwork of colored forms, it gives to the theories of form perception and color-patch perception a special status. They are, as it were, propaedeutic to the theory of visual perception.

Second is the idea that one should examine the retinal image for copies of environmental aspects, and where copies are not found, those aspects are said to be inferred, guessed at, or created. For instance, the third dimension does not have a copy in the retinal image and so it must be constructed; similarly the shape of an object (its structure in 3-space) is not replicated in the image and so it must be constructed or inferred from the two-dimensional outline of the object, which is replicated in the image; and by the same token the arrangement of flashes for the phi phenomenon does not produce a retinal copy of physical

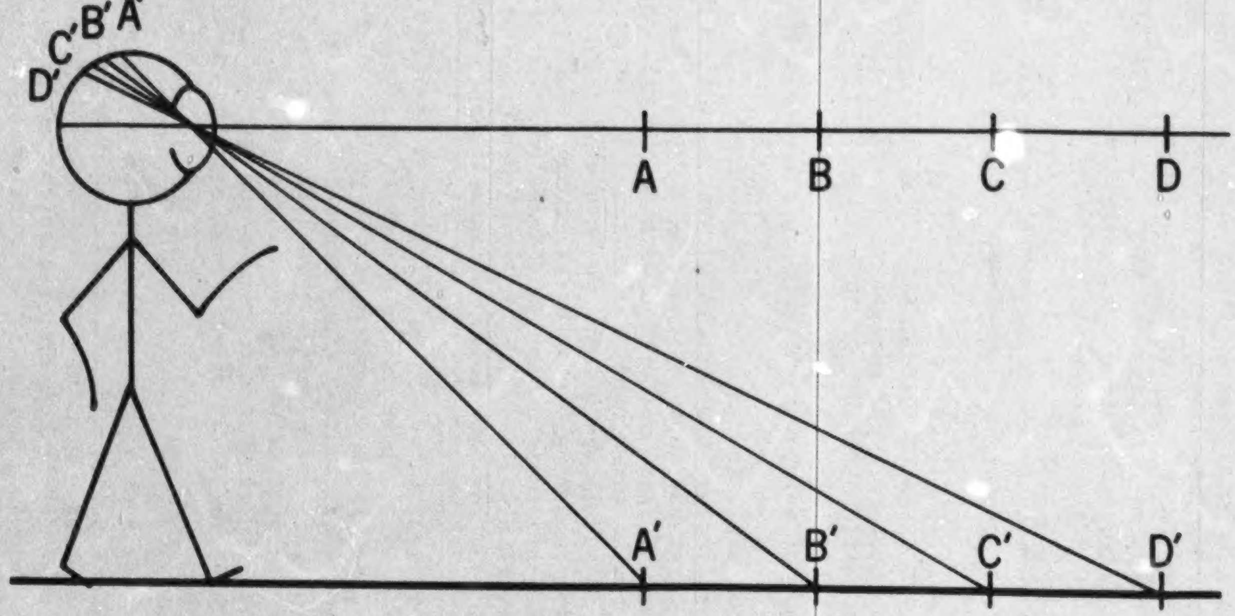

Figure 6: The traditional conception is of an eye registering points in space (ABCD). Gibson's conception is of an eye attached to a body registering points on a surface (A'B'C'D').

FIGURE 6

movement defined as object displacement in space--so the impression of movement in this situation must be a mental creation.

Gibson's (1950, 1966) contention is that this way of conceptualizing visual perception, that is, along the lines suggested in the classical theory of space perception, is blatantly in error. Among his reasons for this contention, the following bear directly on the classical position. For Gibson there is no such problem as the problem of depth or space perception. The very concept of "space" as empty Euclidean space is irrelevant to discourse on perception, and its use in psychology, he argues, is confused and confusing. Animals, he will tell us repeatedly, do not perceive space. We have misconstrued the problem: what animals perceive is the layout of surfaces. And the significance of this restatement of the problem is that in the light there is information for the perception of surface layout--thus surface layout may not have to be created or guessed at, it can be sensed in the meaning of detected.

If Gibson's point of view is valid, then many concepts that evolved with the attempts to solve the space perception problem, and that were perpetuated by its classical solution, may have to be discarded. Let us therefore proceed to the question of what terrestrial environments look like.

Ecological Optics

The tenor of Gibson's denunciation of classical theory is that evolution did not devise visual systems to operate on a vacuous Euclidean space but to detect and interact with the properties of cluttered environments. The environments of animals and man induce inhomogeneities in reflected light, and it is Gibson's principal, guiding intuition that environmental events structure light in ways that specify their properties. The proof of this intuition rests with "ecological optics" (Gibson, 1961), an enterprise that seeks to determine what is contained in the light: if it can be demonstrated that the light is structured by environmental events in a specific fashion, then it is said that the light contains information for these events. In Gibson's view we should look for tighter correspondences between environmental facts and the light as structured by those facts. But all of the above is by way of a summary, and to be appreciated it requires that we reexamine the conception of light with reference to visual perception.

The terrestrial environment is manufactured from solids, liquids, and gases of various chemical composition. The interfaces between these phases of matter are what we normally refer to as surfaces. Surfaces can be said to have structure--that is to say, they are textured--where the texture (or grain) consists of elements of one kind or another duplicated over the entire surface. The "signature" of a surface is given by the cyclicity or periodicity of these textural elements.

But significantly the terrestrial environment as a whole possesses structure at all levels of size--at one extreme it is structured by mountains; at the other, by pebbles and grass (Gibson, 1966). And equally significantly the structure at one level of size is nested in the structure at the next, higher level of size. Each component of the environment can be said to consist of smaller components; facets are nested within facets, and in the most general of senses, forms are nested within forms. Thus the texture of the individual brick is nested within the texture of the wall as given in the arrangement of bricks. We see that it is considerably more apt to treat the structure of the environment as hierarchic, than as mosaic.

The natural environment receives its illumination from the sun, although our living and working environs are mostly illuminated by man-made sources of radiant light. Not all surfaces are directly illuminated by radiant light; some, occasionally most, are illuminated by the light reflected from other surfaces. Natural surfaces tend to be opaque rather than transparent, which means they reflect light rather than transmit it. More precisely, opaque surfaces reflect some portion of incident light and absorb the remainder as a function of their chemical composition. A significant feature of natural surfaces is that they do not reflect light like a mirror, rather they scatter it. This scatter-reflectance is of two kinds: selective for wavelength and unselective for wavelength. In combination, gradations in these two kinds of reflectance, and complex differences in reflectivity--in scattering--due to a textural difference, mean that each different surface modifies light idiosyncratically.

Surfaces, of course, articulate at various angles to each other and when illuminated reflect light among themselves in multiple fashion. The consequence of this is a "flux of interlocking reflected rays in all directions at all points" (Gibson, 1966:12), and, therefore, the light that impinges on an observer in the transparent medium of an illuminated environment is primarily indirect, reflected light.

We need a term to distinguish this light from the radiant light with which we have become most acquainted through physics. Gibson (1961, 1966) suggests the term "ambient" for radiant light that has been modulated by an arrangement of surfaces, that is, by an environment.

There are several significant distinctions to be drawn between radiant and ambient light. Ambient light is in reference to an environment and radiant light is not. Also the two may be distinguished in that ambient light converges to a point of observation, while radiant light diverges from a source of energy. In an illuminated room, wherever one stands there is a sheaf of rays converging at that point. This is a consequence of the multiple reflection (reverberation) of radiant light from the walls, ceiling, and floor. An illuminated medium therefore is said to be packed with convergence points, or points of observation (Gibson, 1961, 1966).

To the bundle of light rays converging to a point of observation we give a special name--the optic array (Gibson, 1961, 1966). Now because surfaces differ in the amount of incident light falling on them, and in inclination, reflectance, and texture, the optic array at a point of observation will consist of different intensities (and different spectral values) in different directions. In short, the optic array has structure and we may in consequence of the conclusion identify a further distinction between radiant and ambient light: the former has no structure and is simply energy, the latter is structured and is potentially information.

What is implied by the radiant/ambient distinction? Assuming that one does not find the distinction quixotic, the foremost implication is that light as stimulation for the ocular equipment of man, animal, and insect has not been properly described by physics. Radiant light has been lucidly portrayed, but not ambient light; and if the preceding is to be respected, it is ambient light that permits perception of the environment.

However, if, following tradition, we remain true to the persuasion that radiant light is the valid and only starting point for the theory of visual perception, then we must conform to the task of explaining how the richness and variety of the visual experience is derived from intensity and wavelength, the variables of radiant light to which brightness and hue correspond as the basic variables of perception. With radiant light as the starting point, we are virtually condemned to a theory of indirect perception

What befalls us when we take ambient light as the point of departure? To begin with, we lose the security of physics, for the stimulus variables of an optic array are not readily couched in the fundamental physical measures. We gain, however, a new perspective on the problem of visual perception, for with ambient light the variables of stimulation are potentially as rich as the variables of experience. Let us dwell, therefore, a little longer on the concept of the optic array.

An especially useful description of the optic array is that it consists of a finite (closed) set of visual solid angles with a common apex at the point of observation and with a transition of intensity separating each solid angle from its topological neighbors (Gibson, 1972). A component solid angle of an array corresponds to a component of the environment--and just as the environment is hierarchically structured, so it is with the optic array: visual solid angles are nested within visual solid angles, and more generally, forms are nested within forms. A few of the variables of the optic array can now be noted: abruptness and amount of intensity transition between visual solid angles; density of intensity transitions within a visual solid angle and change of density across visual solid angles; rates of change, or gradients, in the density of intensity transitions and, with movement, rates of rates of change. This truncated inventory should suffice to suggest that the variables of ecological stimulation are both much richer and of a higher order than those normally adduced for the energy impinging upon an organism.

In summary, explanations of visual perception have traditionally been proposed with reference to the limited variables of radiant light. It would seem, however, that ambient light, that is, radiant light as structured by an environment, is the proper reference. The variables of ambient light are complex and probably limitless. Moreover, there should be a strict correspondence between an ambient optic array and the environment responsible for it. This suggests the following hypothesis: for any isolable property of the environment there should be a corresponding isolable property of the stationary or flowing optic array, however complex. It is this hypothesis, relating to the what of perception, that constitutes the concern of ecological optics. If true, this hypothesis has profound implications for perceptual theory. For if the structured light to the visual system can specify the world and if it is only rarely equivocal (contrary to hundreds of years of studied opinion), then we can imagine that visual systems evolved to be sensitive to its structure, i.e., to the higher-order variables of the optic array. Such being the case, visual perception need not be constructive--it need not be a guessing game--it could be direct. The latter is Gibson's working hypothesis on the how of perception and it may be stated more formally: for every visual perceptual experience, there is a corresponding property of the stationary or flowing optic array, however complex (cf. Gibson, 1959, 1966).

The Concept of Invariance

What then is the nature of the correspondence between the optic array and environmental properties?

Recall that tradition opted for replication as the form of the correspondence between environment and retinal image. One consequence of this has been dealt with; namely, that because the third dimension is not registered in the retinal image, it therefore must be constructed. The upshot of the Gibsonian analysis* is that the correspondence is correlative (in the sense used by analytic projective geometry and not in the sense used by statistics) and that our search should be for correlates of environmental properties, and not for copies. Moreover, one should look for these in the ambient optic array rather than in the retinal image. Where a property of the static or changing optic array corresponds in this sense to a persistent property of the environment, it is referred to as an invariant.

The Stationary and Flowing Optic Array

In the current scheme the natural stimulus for an ocular system is the optic array. This contrasts with the view discussed earlier of stimuli as punctate and momentary, because an optic array by definition is extended and persistent. Furthermore, stimuli are traditionally conceived to be at the retina (hence the primacy of the retinal image), whereas the optic array clearly is not. Indeed, we can mathematically describe the optic array at a position (point of observation) in an illuminated environment without assuming that the position is occupied by an observer (Gibson, 1966).

A significant feature of Gibson's conception of a point of observation is that it is stationary only as a limiting case, simply because observation generally accompanies movement (i.e., the observer is often in motion).

A moving point of observation means a changing optic array: component visual solid angles will be transformed, some will even cease to exist. But the surrounding layout of surfaces responsible for the optic array at a stationary point of observation must also be responsible for the flowing optic array at a moving point of observation. There ought to be abstract optical relations that persist during the course of change, and these relations ought to be specific to the persisting or permanent properties of the environment. On the other hand, the nonpersisting features of the optic array are due to the motion of the point of observation, that is, to locomotion itself. Of the changes in the optic array, we will say that they specify locomotion, and herein lies a significant insight.

Briefly, the gist of the preceding may be conveyed in this fashion: the invariant structure of the changing optic array is information about the environment, that is, it is exterospecific; the variant structure, on the other hand, is information about the observer and therefore may be referred to as propriospecific. We proceed from here with the notion of invariant structure (drawing a comparison between invariant information in the changing and in the unchanging, stationary optic array) leaving to a later moment the egospecific character of variant optical structure.

Consider a problem encountered earlier, that of perceiving continuous distance. From a stationary point of observation, a receding textured surface projects a gradient of optical texture density. (By "gradient" is meant nothing more than an increase or decrease of some measured quantity along a given dimension or axis.) For two identical (but hypothetical) visual solid angles, the one corresponding to a region of the surface close to the point of observation will be less densely packed with intensity transitions (produced by the surface's textural elements scattering the light) than the one corresponding to a region farther away. And, of course, for each identical visual solid angle sandwiched between these two the density of transitions would increase with the distance of the region to which the visual solid angle corresponded.

In the case of a moving point of observation, we can suppose that the impression of distance is given by a dynamic transformation of texture, more precisely, by a gradient of flow velocities: the rate of textural flow for "close by" would be greater than that for "far away." In sum, there are correlates of distance in both the flowing and stationary optic array. The impression, however, is that the flowing optic array is less equivocal in its specification of distance, and of surface variables such as slant (cf. Flock, 1964). But this is only an instance of what is a most general rule, namely, the superiority of changing optic arrays over stationary optic arrays as information about an environment (Gibson, 1966). We should not be surprised by this observation. After all, if a stationary point of observation is a limiting case of a moving point of observation, then it follows that the structure of the optic array at a stationary point is but a special case of the structure of the optic array at a moving point. One can contrive stationary perspectives to fool an observer; to fool a moving observer is an incomparably more difficult task. The notorious Ames room--a distorted room with unpatterned walls and floors--is a case in point. Limited to a stationary perspective, the observer is fooled, but he detects the ruse once he is allowed to move.

It is already understood that the ambient optic array as the natural stimulation for vision has structure: it has some degree of adjacent order and some degree of successive order. What we now understand is that, concurrently, it has components of change and nonchange. For Gibson the import of change is that it reveals what is essential, namely, invariance. We can put this another way: structural variation reveals structural nonvariation. In this conception, movement on the part of the organism serves a singularly important geometric function. There is structure that can be detected in a frozen array, but there is considerably more structure to be detected in a changing array, and the organism can effect the change. In short, the movements of the observer transform the ambient optic array and thereby enhance the detection of invariants.

The Visual Perceptual System

The above reasoning motivates the conception of the mechanism of vision as an actively exploring system. Traditionally, however, a sense system is conceived more conservatively as a well-specified anatomical route that delivers sensations to identifiable centers; integration and cross correlation of sense data then follow as circumstance demands. But in an ecological perspective, the system responsible for seeing ought to be described in ascending levels of activity; for example, mobile eye-mobile eyes in mobile head-mobile eyes in a mobile head on a mobile body. A hierarchy of levels of activity constitutes the visual perceptual system, with each level of activity, from lesser to greater, permitting increasingly more complex transformations of the optic array and, concomitantly, the detection of increasingly higher-order invariants. This conception, of course, does not deny that an eye is for light, but it does question the assumption that seeing is entirely an act of the brain and can be understood solely as such. Seeing, Gibson would say, is an act of the organism.

Consider if you will what transpires when an observer is confronted with an interesting but imperfectly clear object at some distance from her. What does she do to increase the clarity of her perception? She could, of course, adjust the lenses of the eyes, but an eminently more significant improvement would follow upon her walking up to the object. The moral is simple. Locomotion is not immaterial to vision and an eye is all the better for having legs.

It is instructive in this context to consider what is meant by visual information processing. From the traditional interpretation of vision as a sense system, the meaning is clear: there are signals emanating from the environment yielding neural signals to be interpreted. But for Gibson, processing cannot mean this. Quite to the contrary, processing resides in the activities of the perceptual system, in its adjusting, orienting, exploring, and optimizing of information about the environment that is external to the perceiver (Gibson, 1973). On this view, processing is not in the brain as such but pervades the continual cybernetic loop of afference-efference-reafference that defines the synergistic relation between the perceiver and his ecology.

Equivalence of Perceptual Systems

Let us now return to the characteristics of the optic array in order that we may touch upon a further aspect of the concept of invariance with respect to the notion of a perceptual system. We commented above that the optic array as the natural stimulation for the visual perceptual system consisted of order and change of order. What we now recognize is that these characteristics are not modality specific and may be taken as intrinsic to stimulation for any perceptual system. This being the case, one may conjecture that the same structure, and thus the same environmental event, can be equally available to more than one perceptual system. For example, the pattern of discontinuity formed by rubbing or scraping the skin with an object might have the same abstract description as the pattern of acoustic discontinuity in the air determined by rubbing and scraping the same object against other objects.

Experiments into seeing with the skin speak to this very issue (White, Saunders, Scadden, Bach-Y-Rita, and Collins, 1970). The "observer" in these experiments is attached to a system that uses a television camera as its eye, where the camera is quite mobile and can be aimed variously by the observer. The video image is electronically transferred and delivered to a 20 × 20 matrix of solenoid vibrators that stimulate an approximately 10-in.-square area on the observer's back. Each solenoid vibrates when its locus is within an illuminated region of the camera field; thus the matrix as a whole yields a pattern of discontinuity on the skin that corresponds essentially to the arrangement of intensity transitions in the optic array.

Of the many intriguing results reported, the following are the most significant to our concerns. First, properties of the environment are specified in the tactile array. Second, there is no qualitative difference between blind observers and blindfolded normal seeing observers; nor are there any quantitative differences of any great significance. Both blind and sighted (but blindfolded for testing) observers learn to detect the environmental properties specified in the tactile array with virtually equal facility. Third, the tactile array specifies the layout of surfaces and the whereabouts and identity of objects significantly better when it is changing than when it is static. In brief, the observer's ability to detect environmental facts is enhanced when he can move the camera, that is, when he can transform the tactile array. As with the optic array, variation in the tactile array reveals nonvariation; it permits the isolation of what is essential from what is not.

That a changing tactile array can yield information about the three-dimensional structure of an environment--more precisely, the surface layout--and that its pickup does not require prodigious training (see White et al., 1970) is most favorable to Gibson's position and embarrassing to the traditional one. The fact that the tactile array at a moving point of observation contains the same adjacent and successive order--the same abstract mathematical structure--as is contained in the transforming optic array means that the same invariants are there to be detected. On this account, the tactile stimulation does not have to be cross-correlated with information detected by other perceptual systems, or compared with memorial information, in order to be rationalized--to the contrary, the tactile arrangement affords the same meaning as the optical arrangement.

At this juncture it is appropriate to draw a further contrast between the constructivist perspective, in particular the information-processing approach, and the Gibsonian. The emphasis on transforming arrays in the preceding discussion implies that perception is easier, in some sense of the word, when the input variation is greater. But common to a great deal of information-processing modeling is the principle that the nervous system is severely limited in its information-handling capabilities; the implicit assumption is that the fewer the sensory events to be handled per unit of time, the more proficient the nervous system is as an information-processing device. The contrast between the two perspectives, the information-processing/constructivist and the Gibsonian, is highlighted in the following comments by White et al. (1970:27):

> Visual perception thrives when it is flooded with information, when there is a whole page of prose before the eye, or a whole image of the environment; it falters when the input is diminished, when it is forced to read one word at a time, or when it must look at the world through a mailing tube.... The perceptual systems of living organisms are the most remarkable information-reduction machines known. They are not seriously embarrassed in situations where an enormous proportion of the input must be filtered out or ignored, but they are invariably handicapped when the input is dramatically curtailed or artificially encoded. Some of the controversy about the necessity of preprocessing sensory information stems from disappointment in the rates at which human beings can cope with discrete sensory events. It is possible that such evidence of overload reflects more of an inappropriate display than a limitation of the perceiver.

Let us consider one further example of the equivalence of perceptual systems. Consider an object on a collision course with an observer. Optical concomitants of this event consist, in part, of a symmetrically expanding radial flow field kinetically defined over the texture bounded by the object's contours and, simultaneously, the occlusion and disocclusion of texture at the contour edges. If this mathematically complex change or some reasonable facsimile of it is simulated on a screen, the observer will involuntarily duck or dodge (Schiff, Caviness, and Gibson, 1962; Schiff, 1965). And he will do so whether human (adult or infant) or animal (e.g., monkey, chicken, crab).

For simplicity, we may say that the acceleration in symmetrical expansion, referred to as "looming" (Schiff et al., 1962), specifies impending collision. Now it is the case that the participants in the seeing-with-the-skin experiments would often give startled ducks of the head when the part of the tactile array corresponding to an object was suddenly magnified by a quick turn of the zoom lever on the camera (White et al., 1970). One anticipates the Gibsonian interpretation of this curious fact: the mathematical description of the changing optic array and of the changing tactile array corresponding to looming are identical. That is to say, that even though there are obvious qualitative differences between visual and tactual experience, both perceptual systems detect the same information specifying the event of looming.

And consider now the acoustic array as structured by a car on a collision course with you. Again we will have to entertain the hypothesis that the mathematical description of this looming is the same as that for vision and touch. In sum, perceptual systems as different modes of attention sample the same world, and when the same event is detected, it is because of an abstract identity in the available structure to which each system is sensitive. Looming as an environmental event structures the optic array, the tactile array, and the acoustic array in the same fashion.

But what then becomes of the notion of sensation in this scheme of things? For Gibson (e.g., 1966), sensation is largely divorced from perception, in contrast to the more traditional orientation that classifies sensation as the necessary precursor to perceptual experience. Gibson's stance on this point is very similar to that of Heider (1959), who distinguishes between thing (message?) and medium. The human perceiver may attend to either. If he attends to the way in which the medium is structured by the environment, then he perceives; if, on the other hand, he attends to the medium itself, then he is said to have sensations. Given our current understanding of infant vision (cf. Bower, 1971), which suggests that infants see objects in definite ways and at determinate distances, we might be led to conjecture that attending to the way in which light is structured by the environment--that is, attending to the message--is ontogenetically prior to the ability to attend to the medium itself. On this conjecture, sensitivity to the medium and interpreting the medium reflect a latter-day sophistication, whereas at the outset the visual perceptual system is geared to detecting "messages."

The reader may be of the opinion that I am belaboring a somewhat arbitrary and even meaningless distinction. Or, more simply, perhaps a discussion of the sensation/perception distinction seems out of place in a treatment of current thought on visual perception. For myself, however, I am convinced that much of current theory acknowledges implicitly the view that sensation is prior to perception. As hinted earlier, constructivism has departed little from the structuralist notion of discrete sensations as the sum of what is directly accessible. Fundamentally, all that has been done is to substitute the term "feature" for "sensation." And whether antiquated sensations or modern features are taken as directly given (with all else as inference or construction) Ryle's (1949) prisoner-in-the-cell parody is still fashionable (cf. Turvey, 1974). As the parody goes, a prisoner has been held in solitary confinement since birth. His cell is devoid of windows but cracks in the walls provide occasional flickers of light, and through the stones occasional scratchings and tappings can be heard. On the basis of these snippets of light and sound, our hapless prisoner apprehends the unobserved happenings and scenes outside his cell such as football games, the Miss America Pageant, and audiences at the World Congress on Dyslexia. But we should ask, as Ryle does, how could our prisoner ever come to know anything about, say, football games, except by having perceived one in the first place?

Information about One's Self

We recall the conclusion drawn about the flowing optic array at a moving point of observation: the change is propriospecific, in contrast to the nonchange, which is exterospecific. Having deliberated on the contrast between perceptual systems and information about the environment, on the one hand, and sense systems and sensations on the other, we are now in a better position to appreciate the implications of the egospecific nature of vision.

On the classical Sherringtonian view (Sherrington, 1906), each sense or receptor system performs a unique function--it is said to be exteroceptive, proprioceptive, or interoceptive. This view was buttressed by an older and more sacred doctrine, namely, Johannes Müller's law of specific nerve energies, which holds that sensation is specific to the receptor initiating it. Müller's doctrine argues that one's awareness is of the state of the nerves; thus, by implication, an awareness of movements of the self is an awareness of the states of a specialized receptor system. Tradition referred to this awareness as proprioception/kinesthesis and localized the specialized receptor system in the muscles and joints.

Gibson's rebuttal to these conceptions is by now quite evident. If by exteroceptive we mean detecting information about events extrinsic to the organism and if by proprioceptive we mean detecting information about the animal's own bodily activities, then clearly vision is as capable of the latter as it is of the former. The classical dichotomy of exteroceptive and proprioceptive systems is wrong. Thus we may speak meaningfully of "visual proprioception" in reference to the visual perception of bodily movement (Gibson, 1966; Lishman and Lee, 1973; Lee and Aronson, 1974) to emphasize that proprioception is not the prerogative of a specialized receptor system or sense organ. Indeed, one can make an argument that visual proprioception is the more reliable, and often the only reliable, source of information about movements of the self (egomovement). For example, muscle-joint kinesthesis is uninformative when I am traversing an environment by car or by train. But the flowing optic array would appear to specify locomotion of the observer, be he passive or active. And what of the fish swimming upstream or the bird flying against a head wind? Muscle-joint kinesthesis would specify movement in the sense of change of location where there is none.

For a locomoting observer the flow pattern ought to provide information for getting about in an environment, that is, for guiding or steering one's movements. The principal feature of the flowing optic array concomitant to locomotion is that it is a total transformation of the projected environment. The array expands ahead and contracts behind. The projection of the place in the environment to which the observer is heading is the focus of expansion in the ambient optic array, while the focus of contraction is the direction from which the observer has just come. Thus there is information about the direction of movement, and to steer toward a particular object means essentially to keep that object--as a closed internally textured contour in the optic array--as the focus of optical expansion (Gibson, 1958; Johnston, White, and Cumming, 1973). In a detailed mathematical analysis of the transforming optic array at a moving point of observation, Lee (1974) develops a proof of the availability of body-scaled information about the size and shape of potential obstacles to locomotion. Relatedly, the claim can be made that there is body-scaled information relevant to the amount of thrust needed to leap over an obstacle and to the time at which the thrust is to be applied. In sum, for an animal capable of registering the optical velocity field and its derivative properties, there is directly specifiable optical information for the guidance of locomotion in a cluttered environment.

But do I need to view the focus of expansion in order to know in which direction I am moving? We know, of course, that we can perceive our direction of movement when the head is turned sideways. Is it possible therefore that samples of the total optical flow pattern, samples that do not contain the focus of expansion, can specify the direction of one's motion? The answer is provided through the use of computer-generated motion pictures depicting various flow patterns for egomotion over an endless textured plain (Warren, 1975). Samples of optical flow patterns can be generated that give specific impressions of movement toward specific locations not contained in the samples. Aside from highlighting the broad egomovement specificity of the changing optic array, a particularly significant feature of this demonstration is that a sample of the total ambient array can specify the total ambient array. Or, to put it somewhat differently, the flow pattern of the field of view specifies characteristics of the ambient optic array outside the field of view. On this point, Warren (1975) reminds us of Merleau-Ponty's (1947:277), claim that: "We see as far as our hold on things extends, far beyond the zone of clear vision, and even behind us. When we reach the limits of the visual field we do not pass from vision to non-vision...."

The proprioceptive role played by the visual perceptual system enjoys further elaboration in the demonstration that the human infant's ability to maintain a stable posture is very much under visual control. More generally, it is believed that the principal information about body sway is broadcast by receptors in the vestibular canals and in the muscle/joint complexes--mechanoreceptors, as they are often called. However, it can be shown that balance is perturbed by transformations of the ambient optic array in a direction that is specific to the transformations (Lee and Aronson, 1974). From the foregoing account, an expanding optic array specifies forward egomovement, a contracting optic array specifies backward egomovement. An infant standing on a stationary floor in a room can be caused to sway forward or fall forward when the walls and ceiling of the room move away from him (i.e., contraction of the optic array) and to sway backward or fall backward when the "room" moves toward him (expansion of the optic array). Though not as pronounced, the same relation between standing and optical change can be observed in adults (see Lee and Aronson, 1974).

Object-Structure Correlates in the Optic Array and Fourier Analysis

To a very large extent the preceding sections have assumed a "frozen" environment. Our analysis has concentrated on the stationary and the moving point of observation in respect to an illuminated environment, which was described tersely as a fixed arrangement of textured surfaces. Of course, implicitly it was assumed that the environment was cluttered with objects--themselves arrangements of surfaces--but little comment was made about them or about their activities, with the exception of a passing reference to "looming." In the sections that follow, this omission is partially remedied as we direct our attention, in measured steps, to objects and the changes they undergo. In what follows, we unfreeze the environment.

But first a brief consideration of the information about a nontransforming object at a stationary point of observation is in order. (The object in mind is one that is detachable from the ground plane: it is movable.) Expressed roughlv: an object's concomitant in the optic array is a visual solid angle, corresponding to the exposed face(s) of the object, packed with smaller (and nested) visual solid angles, corresponding to the facets or grain of the exposed face(s). One might say for simplicity that the concomitant is a closed contour with internal texture.

We can raise two elementary but instructive questions. First, what kind of ordered discontinuity in the arrangement of ambient light separates the visual solid angle corresponding to an object from the other visual solid angles in which it is nested? More precisely, what is the mathematical invariant for contour? Second, what kind of ordered discontinuity in the optic array is specific to the type of articulation between surfaces comprising an object? More bluntly, what is the mathematical invariant for edge type?

We are reminded that both of these questions were raised earlier in a different way and in a different context, that of seeing machines, where the concern is with line drawings of variously arranged polyhedral bodies and how a three-dimensional description of them can be recovered. It is important in that connection to devise heuristics for separating bodies from bodies and for identifying the quality of the joint between surfaces. In the current context we are assuming a natural environment, as opposed to a line drawing, and ecological optics, as opposed to image optics. Our current intuition is that the light to an eye should specify unequivocally the presence of contours and the type of edge. The following is meant only as a hint of how this intuition might be substantiated.

Solid objects are composed of textured surfaces whose gradients relative to some stationary point of observation are correlates of their shape and slant. For example, as the slant of a surface varies in continuous or stepwise fashion, then the projected gradient ought to do likewise (Gibson, 1950, 1966). In the instance of an object-surface bending or of two surfaces of a polyhedral object joining, the abrupt change in slant will yield an abrupt change in the gradients of optical texture density. Precisely, a concave bend or joint (edge) is specified by an abrupt transition from one rate of change in intensity transitions to a slower rate; a transition tc a faster rate specifies a convex bend or join. These mathematical discontinuities exemplify the form of invariant for edge type. An invariant for contour can be a more simple mathematical discontinuity--a sudden transition in optical texture density.

The above is but a rough account of the optical specification of edge and contour in a frozen array (for a fuller account dealing with transforming arrays, see Mace, 1974). It is instructive though in the following senses: first, it highlights the principle that the optical structure uniquely specifying environmental properties is "...defined not over parameters of radiant light rays such as intensity, but over relations of change in intensity" (Mace, 1974:144); second, it expresses the mathematically abstract nature of invariants--an edge type is given by a particular change in rates of change; third, it strongly implies that the perception of object relations in a scene is predicated upon optical variables and mechanisms quite different from those currently envisioned by artificial intelligence research.

The last point is well worth developing. In most scene analysis programs there has to be a method for mapping lines in the picture domain onto edges in the scene domain. But in the Gibsonian view, lines are not primitive entities from which all else is constructed. The organism is confronted by mathematical relations in the light, relations across changes in intensity transitions, which are specific to edge types. If he can pick up on these higher-order relations, he has no need for lines nor for line-to-edge mapping rules.

Mackworth (1973) has described a computer program for interpreting line drawings of polyhedral scenes that does not rely on knowledge of specific object prototypes (cf. Falk, 1972) nor does it appeal to the technique of mapping vertices onto corners (cf. Clowes, 1971). The kernel feature of Mackworth's program is the mapping of surfaces into a gradient space and the use of this representation to determine (through the use of "coherence" rules that articulating surfaces must satisfy) the properties of the scene. Of course, Mackworth's program is designed for line drawings (i.e., textureless regions) and therefore requires the use of heuristics to determine the pattern of surface orientations. But one is motivated by this program to ask how computationally more straightforward scene analysis programs would be if they worked on natural scenes instead of line drawings, and were equipped with the means to determine density and discontinuities of texture. For a long time Gibson has argued that picture perception is not a simpler version of natural scene perception. His suspicion is that, on the contrary, an account of the latter is more readily forthcoming than an account of the former.

The tenor of the immediately preceding comments suggests that it would be useful to grasp, if only in a rough and approximate way, the kind of mechanism permitting the detection of textural variables. Understandably, little effort has been directed to the determination of how surface structure is perceived because most investigators interested in the how of perception are firmly entrenched in retinal image optics. Consequently, they have been satisfied with inquiries into two-dimensional outline forms and features. However, research in computer vision has not been so remiss (see Hawkins, 1970) and one recent report is especially interesting: Bajcsy (1973) points to the special advantages of surface texture descriptions derived in the Fourier domain. Why this should be thought interesting becomes evident when one considers a rather peculiar version of the cross-adaptation experiment described earlier.

Consider a grating of vertical dark bars of equal width on a light background such that the plot of luminance against spatial location yields a regularly repeating square wave. Now consider a further grating in which the relation between bars and background yields a sinusoidal plot of luminance against spatial position. Most clearly we could generate an indefinitely large number of gratings yielding square waves and sinusoids of different spatial frequencies as measured in cycles per degree of visual angle. The use of gratings such as these in cross-adaptation (and related) experiments suggests that the visual system is selectively sensitive to spatial frequency--that it performs a spatial frequency analysis (Cornsweet, 1970; Sekuler, 1974).

A stronger inference, but one that is not dictated by the data (see Sekuler, 1974), reads as follows: there is a set of independent channels each sensitive to a particular spatial frequency band whose overall behavior may be characterized as that of performing an analysis into Fourier components (e.g., Pollen, Lee, and Taylor, 1971). By way of illustration, Blakemore and Campbell (1969) demonstrated that adapting to a high-contrast square wave grating of frequency F raised the threshold for sinusoidal gratings of frequency F and (to a lesser degree) sinusoidal gratings of frequency 3F. If the visual system is performing a Fourier analysis, then the square wave grating ought to be "decomposed" into its fundamental and odd harmonics. And, consequently, there ought to be a change in the sensitivity of the system to sinusoidal gratings at the fundamental and odd harmonic frequencies. In this respect, the Blakemore and Campbell finding fits the Fourier model. A further fit is a demonstration by Maffei and Fiorentini (1972) of Fourier synthesis: a perceptual impression of a square wave grating of frequency F results when two sinusoidal gratings, one of frequency F and one of frequency 3F, are presented separately, one to one eye and one to the other.

Though discoveries like the above have aroused considerable interest and are proliferating at an enormous rate (see Sekuler, 1974), investigators have yet to concern themselves seriously with the role played by frequency analysis in visual perception. But surely Gibson's (1961, 1966) ecological analysis and Bajcsy's (1973) insight suggest the role: natural environments consist of surfaces; natural surfaces are textured, in the sense of repeating patterns of varying spatial frequency; and the detection of surface properties, static and kinetic, is mandatory for a being that must adapt to its world.

Event Perception: Structural and Transformational Invariants

Let us now come to an examination of objects undergoing change. The examination remains in relation to a stationary point of observation.

Shaw et al. (1974:279) have remarked: "The environment of any organism is in dynamic process such that the smallest significant unit of ecological analysis must be an event rather than a simple stimulus, object, relation, geometric configuration or any other construct whose essence can be captured in static terms." "Event" is the term most befitting a change in an arrangement of surfaces; and rolling, opening, falling, rotating, and aging are examples of events. Nonchange is but a special case of change; consequently, resting, standing, being supported, etc., are legitimate events. However, the concept of an event cannot be meaningfully construed in the absence of a subject; thus events are more accurately exemplified by: "the ball rolls," "the door opens," "the rock falls," and "Bill grows old." The last is particularly instructive because it reminds us that the events we perceive can be of varying duration--the "slow" event of Bill's aging contrasts with the "fast" event of rocks falling.

In concert with the fundamental hypothesis of ecological optics, the light is structured by an event in a fashion specific to that event. An event is said to modulate the light in a way that specifies the identity of the participant in the event and the dynamic component of the event, the form of the change. Consequently, for event perception it is conjectured that the visual perceptual system must detect the structural invariant specifying the structure undergoing the change and the transformational invariant specifying the nature of the change undergone (Shaw and Wilson, in press).

Let me proceed to elaborate on these conjectures. A demonstration by Gibson and Gibson (1957) provides an appropriate example. The observer viewed a screen onto which was projected the shadow of one of the following: a regular form (a solid square), an irregular form (ameboid shape), a regular pattern (a square group of dark squares), an irregular pattern (an ameboid group of ameboid dark shapes or spots). Each silhouette was semirotated cyclically and the observer had to report on the degree of change in slant. For both the regular and irregular silhouettes, the observer's experience was of a rigid constant surface changing slant and he was able to judge the slant most accurately. The implication of this simple demonstration is quite paradoxical as Gibson and Gibson (1957) realized: a change of form yields a constant form together with a change of slant. The paradox originates, in part, in the two uses of the term "form." When we speak of "change of form," we are referring to the abstract geometrical projected form, or silhoustte; when we refer to "constant form," we are speaking about the rigid substantial form. The traditional way to interpret kinetic depth effects, such as that just described, is to say that the currently perceived static and flat projected form is combined with the memories of the preceding static and flat projected forms to yield, through an act of construction, the substantial form-in-depth (cf. Wallach and O'Connell, 1953). And it is evident that this interpretation is motivated by the two assumptions of (1) stimuli as momentary frozen slices in time and (2) retinal image optics as the proper departure point for speculation. Quite to the contrary is the Gibsonian interpretation that follows from ecological optics with its notion of an enduring ambient optic array: a lawful transformation of the optic array specifies both an unchanging rigid object and its motion. The emphasis here is on the event's modulation of the structure of the ambient light.

The implication of this interpretation for the understanding of object-shape perception is quite radical and exceedingly difficult to grasp on initial reading. In any event let me take the liberty of spelling it out relying on subsequent discussion for its clarification. The shape of an object is not given by a set of frozen perspectives but by a unique transformation of the optic array. Or, synonymous, object-shape perception is not based on the perception of static forms but on the perception of formless invariants detected over time (cf. Shaw et al., 1974). This hypothesis denies that shape can be captured by the formal descriptions of classical geometry; that static forms are unique entities (to the contrary, a form is not a thing but a variable of a thing); that static forms are primary sources of data for perception; and that shape is an isolable physical property of objects. On the latter point we may comment that shape is perhaps better conceived as a property of events (Shaw et al., 1974; Shaw and Wilson, in press). Before further examination of this hypothesis let us consider an example of a transformational invariant.

The sensibility of the observer to the form of change in a changing optic array is demonstrated in a singularly elegant fashion by von Fieandt and Gibson (1959). Consider the contrast between rigid and nonrigid (elastic) motions. An observer peers at a screen onto which is projected the shadow of an irregular elastic fishnet. The fishnet is attached to a frame that is manipulable in the following ways: one end of the frame can be slid in and out, compressing and stretching the shadow (an elastic transform), or the whole frame can be semi-rotated back and forth subjecting the shadow to foreshortening and its inverse (a rigid translation). All the observer witnesses on the screen is the changing textured shadow, that is, he does not see the shadow of the frame, only the changing motion of the textural elements, which are virtually quantitatively identical for the two transformations. Yet the observer readily perceives a distinct elastic (topological) transformation, in the one case, and a distinct rigid transformation, in the other. In short, he is sensitive to the quality of change, i.e., to the transformational invariants in the ambient optic array that specify elastic and rigid happenings.

Now the explicit and implicit strands of the currently developing thesis can be woven together. At a stationary point of observation, a detached object, as a surface arrangement, has specifiable structural concomitants in the ambient optic array. If the object is caused to move, or if it is squashed, or if it grows, or if it disintegrates, the ambient optic array will be transformed; that is to say, its structure will be disturbed. The onset and offset of this transformation or disturbance will be identical to the beginning and termination of the event and since the duration of ecologically significant events ranges from milliseconds to scores of years, so it is with the event-related optical disturbance. Though the form of the disturbance does not copy the event, it does correspond mathematically to the event and is said to do so in two ways: it corresponds to the kind of change (transformational invariant) and to the structure whose identity persists in the course of change (structural invariant). Visual perceptual systems are assumed to sense (detect) these invariants.

Objects that participate in events are said to have shape, but it is argued here that shape is more accurately a property of the events into which the objects enter as participants than of the objects themselves. Shape, we have said, is a formless invariant over time and to this proposition we now turn.

Symmetry Groups and Shape Perception

Intuitively a circle and square both have symmetry, but they do not have the same symmetry. How might we describe the idea of symmetry so that we can capture this difference? The mathematician's answer is that symmetry is concerned with certain rigid mappings (referred to as symmetry operations or automorphisms) that leave sets of points unchanged. Thus mathematically any rotation in the plane about its center leaves the circle, as a set of points, invariant. On the other hand, only certain rotations (i.e., the integer multiples of $\pi/2$) map the square to itself. Evidently the properties of circle and square could be revealed and contrasted in the symmetry operations that leave them unchanged. Moreover, we can recognize the synonymity of the concepts of symmetry and invariance: paraphrasing Weyl (1952), a thing is symmetrical if there is anything we can do to it so that after we have done it, it appears the same as it did before (cf. Feynmann, 1967; Shaw et al., 1974). For the mathematician, those symmetry operations that leave an object unchanged form a group, and this group describes exactly the symmetry (invariance) possessed by the object. Though we cannot draw an exact parallel between the mathematician's notion of symmetry and Gibson's notion of invariant, we can use the mathematician's insights as a guideline. Ideally, the reader will find this particular guideline useful for understanding the ideas expressed in the last section.

To illustrate, consider the symmetry of a square. There are eight operations that leave the square invariant (i.e., map the square into itself): the null operation (no change), the clockwise rotations through $\pi/2$, π, and $3\pi/2$; and the reflections in the horizontal axis, vertical axis, and the two diagonal axes. Now the composition of any two symmetry operations is itself a symmetry operation, thus reflection in the horizontal axis and the clockwise rotation by π is the same as reflecting in the vertical axis. Thus a table can be constructed of the composition of each symmetry operation with each other symmetry operation. This table, or more accurately the mathematical structure of this table, describes the symmetry of the square, and it has certain interesting properties possessed by all symmetry tables.

1. If *a* and *b* are symmetry operations, then *a o b* is a symmetry operation, where *o* is some principle of combination. This is the property of closure.

2. If *a*, *b*, and *c* are symmetry operations, then (*a o b*) *o c* = *a o* (*b o c*). This is the property of associativity.

3. There is a symmetry operation *e* such that *a o e* = *a* = *e o a*. Here *e* is the identity symmetry operation.

4. Given any symmetry operation *a*, we can always find a symmetry operation *b* such that *a o b* = *e*. For every operation there is an *inverse*.

These four properties define what is referred to mathematically as a group.

We can now return to our hypothesis on shape perception remarking after Shaw et al. (1974) that rigid shape is the structural invariant of an event whose transformational invariant is formally equivalent to a symmetry group, that of rotations. A partial substantiation of this notion is given in a demonstration in which a wire cube is rotated at constant speed on each of its axes of symmetry and strobed at appropriate rates (Shaw et al., 1974). When rotated on a face, self-congruence is achieved every 90°; when rotated on a vertex, positions of self-congruence occur every 120°. It follows that the group description of the cube's symmetry will differ as a function of the axis of rotation.

For rotation on a face, the period of symmetry is four; for rotation on a vertex, the symmetry period is three. In the demonstration, the strobe rate is either synchronous, or asynchronous with the period of symmetry (that is, the strobe rate is either an integer multiple or not an integer multiple of the symmetry period). When the strobe rate is synchronized with the symmetry period, a cube is seen, when the strobe rate is not synchronized with the symmetry period, the shape of the object is no longer recognizable as "cube." And significantly a strobe rate that permits the perception of cube shape when the cube is rotating on a face does not permit that perception when rotation is on a vertex, and vice versa.

This demonstration is significant in the following respects. First, it is evident that shape is not specified by any arbitrarily chosen set of variant perspectives. On the contrary, it appears that rotational symmetry must be preserved in the variant perspectives projected to the eye. Asynchronous strobing annihilates this symmetry and results in a set of ordered perspectives that specify some other shape. In short, different successive orderings of perspectives specify different shapes of the same physical object and we can now understand what it means to say that shape is a property of events rather than of objects. Second, it is clear that not all perspectives are needed to specify shape, only a special ordered subset of those perspectives. There is reason to believe that the "special ordered subset" may be well-defined mathematically; Shaw and Wilson (in press) conjecture that it is the generator set of the group. All in all, there is some justification for Gibson's (1950, 1966) claim that shape perception is based on the perception of formless invariants detected over time.

We can further understand the conception of formless invariants by considering an event which, quite unlike the rotating cube, can be said to transpire slowly.

Of some considerable significance to our everyday living is the ability to determine the relative age-level of faces. All things considered, we manifest this ability rather well. In the light of our current discussion of events we might venture to say that this ability reflects our sensitivity to a particular class of symmetry operations--those that characterize aging. For most assuredly when we speak of the face aging we are speaking of a remodeling of the head that leaves invariant the structural information specifying the species and the more specialized structural information affording recognition of the individual.

The biologist D'Arcy Thompson (1917) had the insight that transformations of a system of spatial coordinates permit the characterization of the remodeling of plants and animals by evolution. The particular advantage of the method of coordinate transformation lies in the fact that one can uncover the appropriate symmetry operations in the absence of a complete mathematical description of the object to be transformed. Pittenger and Shaw (in press) applied this insight to aging. They inscribed a profile in a space of two coordinates and then subjected the coordinate space to the affine transformations of strain (a transformation that maps a square into a rectangle) and shear (a transformation that maps a square into a rhomboid). These symmetry operations--for with appropriately chosen parameters, they preserve the identity of the individual and the species--suitably described the transformational invariant of aging: the original profile could be mapped into younger and older versions whose relative age levels could be accurately ranked by observers. Of the two transformations, strain was the more significant.

The argument emerges that the perception of aging is not necessarily based on the discrimination of local features but rather on the perception of global invariant information of a higher order that is detected over time. Indicative of the higher-order nature or abstractness of this information is the fact that when the strain transformation is applied to the profile of an inanimate object, such as a Volkswagen "Beetle," it generates a family of profiles that can be rank ordered for age commensurate with the rank ordering of faces (Shaw and Pittenger, in press). The implication is that the transformational invariant for aging is independent of the features common to all animate things that grow--just as it is most obviously independent of the features of any single face, or of the features common to all faces.

Indirect and Direct Perception: A Brief Comparison and Summary

The purpose of this paper was to provide an overview of perceptual theory as it relates to vision, together with a liberal sprinkling of empirical facts. To this purpose I have dealt at length with what I take to be the two major and contrasting perspectives: that visual perception is indirect and a derivative of conception; that visual perception is direct and independent of conception. The focus of this final statement is that the distinction between the two perspectives turns on the issue of what order of physical space is the proper basis for the theory of visual perception.

In the main, conjectures on the nature of visual experience derive from a framework that takes as its departure point a bidimensional description of the environment and that seeks to explain how descriptions in three and four dimensions are achieved through mental elaboration. A consensus is that these higher-order accounts are the result of inference and memory, and it is evident how the doctrine of a 2-space world as the sum of what is given dictates an attitude of visual perception as constructed. We have remarked earlier on the centuries old hegemony of this doctrine (see Pastore, 1971).

In opposing the official doctrine, the view of perception as primary takes kinetic events rather than static two-dimensional images as the proper point of departure (see Gibson, 1966; Johansson, 1974; Shaw et al., 1974). More precisely, it is argued that the theory of seeing ought to be anchored in four-dimensional space rather than in the two-dimensional space favored by tradition. We may suppose that for naturally mobile animals and primitive man the instances of pure, static perception are rare. Moreover, "The structuring of light by artifice" (Gibson, 1966:224)--the representation of environments and events by picture--is a relatively recent addition to man's ecology. The argument, in principle, is that the variety of visual perceptual systems evolved by nature are incomparably better suited to the transforming optic array--to the mathematically abstract, optical concomitants of three-dimensional structures transformed over time--than to the static two-dimensional pattern.

On this argument there should be several consequences of reducing the dimensionality of the space in which the visual perceptual system operates. First, the sum total of environmental properties mapped by the ambient optic array is reduced. Second, there is a nontrivial change in the class of invariants--spaces of fewer dimensions are accompanied by invariants of a lower order. Third, perceptual equivocality relates inversely to order of invariant; thus, given the second consequence, equivocal perceptions are more likely when the visual system operates in spaces of fewer dimensions. As one might suspect, these consequences have implications for learning.

It is quite obvious that in both perspectives one must learn to perceive. But the kind of learning implied by the traditional perspective differs radically from that which is implied by the Gibsonian perspective. In the former case, one must know in advance something about the environment in order to perceive it properly; thus one must come tacitly to understand a variety of rules and to register a variety of facts in order to make sense of the inadequate deliverances of his visual system. The alternative, of course, is that one cannot know anything about the environment except as he perceives it, or has perceived it. On the alternative view, we should not treat perceptual learning as a matter of supplementing inadequate data with information drawn from memory. Rather we should see perceptual learning as a matter of differentiating the complex, nested relationships in the dynamically structured medium--of tuning into invariants.

But given the above, we may conjecture that perceptual learning is a function of the dimensionality of the physical space in which the visual perceptual system operates. The frozen array, the limiting case of continuous nontransformation, provides a less than optimal set of conditions for visual perception; the invariants are fewer, more difficult to tune into, and less reliable. Distinguishing bidimensional information poses a more difficult problem for the visual perceptual system than distinguishing information of higher dimensions; learning in the former case should be slower and more devious. On this line of reasoning we may conjecture that reading, the distinguishing of information in the "letter array," is not a task to which the visual perceptual system is especially suited, despite its necessary involvement.

REFERENCES

Arbib, M. A. (1972) The Metaphorical Brain: An Introduction to Cybernetics as Artificial Intelligence and Brain Theory. (New York: John Wiley & Sons).

Averbach, E. and A. S. Coriell. (1961) Short-term memory in vision. Bell Syst. Tech. J. 40, 309-328.

Bajcsy, R. (1973) Computer description of textured surfaces. In Third International Joint Conference on Artificial Intelligence. (Menlo Park, Calif.: Stanford Research Institute).

Baron, J. and I. Thurstone. (1973) An analysis of the word superiority effect. Cog. Psychol. 4, 207-208.

Biederman, I. (1972) Perceiving real-world scenes. Science 177, 77-79.

Biederman, I., A. L. Glass, and E. W. Stacy, Jr. (1973) Searching for objects in real-world scenes. J. Exp. Psychol. 97, 22-27.

Biederman, I., J. C. Rabinowitz, A. L. Glass, and E. W. Stacy, Jr. (1974) On the information extracted from a glance at a scene. J. Exp. Psychol. 103, 597-600.

Blakemore, C. and F. W. Campbell. (1969) On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. J. Physiol. (London) 203, 237-260.

Bower, T. G. R. (1970) Reading by eye. In Basic Studies in Reading, ed. by H. Levin and J. Williams. (New York: Basic Books).

Bower, T. G. R. (1971) The object in the world of the infant. Sci. Amer. 225 (Oct.), 30-38.

Brand, J. (1971) Classification without identification in visual search. Quart. J. Exp. Psychol. 23, 178-186.

Campbell, F. W. and L. Maffei. (1971) The tilt after-effect: A fresh look. Vision Res. 11, 833-840.

Clark, S. E. (1969) Retrieval of color information from the preperceptual storage system. J. Exp. Psychol. 82, 263-266.

Clowes, M. (1971) On seeing things. Artificial Intelligence 2, 79-112.

Coltheart, M. (1971) Visual feature-analyzers and after-effects of tilt and curvature. Psychol. Res. 78, 114-121.

Coltheart, M. (1972) Visual information processing. In New Horizons in Psychology, vol. 2, ed. by P. C. Dodwell. (Harmondsworth, England: Penguin).

Cooper, L. A. and R. N. Shepard. (1973) Chronometric studies of the rotation of mental images. In Visual Information Processing, ed. by W. G. Chase. (New York: Academic Press).

Cornsweet, T. N. (1970) Visual Perception. (New York: Academic Press).

Craik, K. J. W. (1943) The Nature of Explanation. (Cambridge, England: University of Cambridge Press).

Doost, R. and M. T. Turvey. (1971) Iconic memory and central processing capacity. Percept. Psychophys. 9, 269-274.

Falk, G. (1972) Interpretation of imperfect line data as a three-dimensional scene. Artificial Intelligence 3, 101-144.

Favreau, O. E., V. F. Emerson, and M. C. Corballis. (1972) Motion perception: A color-contingent after effect. Science 196, 78-79.

Feynmaun, R. (1967) The Character of Physical Law. (Cambridge, Mass.: MIT Press).

Flock, H. R. (1964) Some conditions sufficient for accurate monocular perceptions of moving surface slants. J. Exp. Psychol. 67, 560-572.

Franks, J. J. and J. D. Bransford. (1971) Abstraction of visual patterns. J. Exp. Psychol. 90, 65-74.

Gibson, J. J. (1950) The Perception of the Visual World. (Boston: Houghton Mifflin).

Gibson, J. J. (1958) Visually controlled locomotion and visual orientation in animals. Brit. J. Psychol. 49, 182-194.

Gibson, J. J. (1959) Perception as a function of stimulation. In Psychology: A Study of a Science, vol. 1, ed. by S. Koch. (New York: McGraw-Hill).

Gibson, J. J. (1961) Ecological optics. Vision Research 1, 253-262.

Gibson, J. J. (1966) The Senses Considered as Perceptual Systems. (Boston: Houghton Mifflin).

Gibson, J. J. (1971) The information available in pictures. Leonardo 4, 27-35.

Gibson, J. J. (1972) On the concept of the "Visual Solid Angle" in an optic array and its history. Unpublished manuscript, Cornell University.

Gibson, J. J. (1973) What is meant by the processing of information? Unpublished manuscript, Cornell University.

Gibson, J. J. and E. J. Gibson. (1957) Continuous perspective transformations and the perception of rigid motion. J. Exp. Psychol. 54, 129-138.

Gibson, J. J. and M. Radner. (1937) Adaptation, after-effect, and contrast in the perception of tilted lines. I. Quantitative studies. J. Exp. Psychol. 20, 453-467.

Goodman, N. (1968) Languages of Art: An Approach to a Theory of Symbols. (Indianapolis: Bobbs-Merrill).

Gordon, I. E. and S. Hayward. (1973) Second-order isomorphism of internal representations of familiar faces. Percept. Psychophys. 14, 334-336.

Gregory, R. L. (1969) On how so little information controls so much behavior. In Towards a Theoretical Biology, vol. 2, ed. by C. H. Waddington. (Chicago: Aldine Publishing Co.).

Gregory, R. L. (1970) The Intelligent Eye. (New York: McGraw-Hill).

Gregory, R. L. (1972) Seeing as thinking: An active theory of perception. London Times Literary Supplement. June 23, 707-708.

Guzman, A. (1969) Decomposition of a visual scene into three-dimensional bodies. In *Automatic Interpretation and Classification of Images*, ed. by A. Grasselli. (New York: Academic Press).

Haber, R. N. (1969) Information processing analyses of visual perception: An introduction. In *Information Processing Approaches to Visual Perception*, ed. by R. N. Haber. (New York: Holt, Rinehart & Winston).

Hagen, M. A. (1974) Picture perception: Toward a theoretical model. *Psychol. Bull. 81*, 471-497.

Harris, C. S. and A. R. Gibson. (1968) Is orientation-specific color adaptation in human vision due to edge detectors, afterimages, or "dipoles?" *Science 162*, 1506-1507.

Hawkins, J. K. (1970) Textural properties for pattern recognition. In *Picture Processing and Psychopictorics*, ed. by B. S. Lipkin. (New York: Academic Press).

Hebb, D. O. (1949) *The Organization of Behavior*. (New York: John Wiley & Sons).

Heider, E. R. and D. C. Oliver. (1972) The structure of the color space in naming and memory for two languages. *Cog. Psychol. 3*, 337-354.

Heider, F. (1959) On perception and event structure and the psychological environment. *Psychol. Iss. 1*, No. 3.

Held, R. and S. R. Shattuck. (1971) Color and edge-sensitive channels in the human visual system: Tuning for orientation. *Science 174*, 314-316.

Helm, C. E. (1964) Multidimensional ratio scaling analysis of perceived color relations. *J. Opt. Soc. Amer. 54*, 256-262.

Helmholtz, H. von. (1925) *Treatise on Psychological Optics*. Translated from the 3rd German edition (1909-1911) and edited by J. P. Southall (Rochester, N. Y.: Optical Society of America).

Henderson, L. A. (1974) A word superiority effect without orthographic assistance. *Quart. J. Exp. Psychol. 26*, 301-311.

Hochberg, J. (1968) In the mind's eye. In *Contemporary Theory and Research in Visual Perception*, ed. by R. N. Haber. (New York: Holt, Rinehart & Winston).

Hochberg, J. (1970) Attention, organization, and consciousness. In *Attention: Contemporary Theory and Analysis*, ed. by D. I. Mostofsky. (New York: Appleton-Century-Crofts).

Hochberg, J. (1974) Higher-order stimuli and inter-response coupling in the perception of the visual world. In *Perception: Essays in Honor of James J. Gibson*, ed. by R. B. MacLeod and H. L. Pick, Jr. (Ithaca, N. Y.: Cornell University Press).

Hubel, D. H. and T. N. Wiesel. (1967) Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat. *J. Neurophysiol. 30*, 1561-1573.

Hubel, D. H. and T. N. Wiesel. (1968) Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. 195*, 215-243.

Ingling, N. (1972) Categorization: A mechanism for rapid information processing. *J. Exp. Psychol. 94*, 239-243.

Ittleson, W. H. (1960) *Visual Space Perception*. (New York: Springer-Verlag).

Johansson, G. (1974) Projective transformations as determining visual space perception. In *Perception: Essays in Honor of James J. Gibson*, ed. by R. B. MacLeod and H. L. Pick, Jr. (Ithaca, N. Y.: Cornell University Press).

Johnston, I. R., G. R. White, and R. W. Cumming. (1973) The role of optical expansion patterns in locomotor control. *Amer. J. Psychol. 86*, 311-324.

Johnston, J. C. and J. L. McClelland. (1974) Perception of letters in words: Seek not and ye shall find. *Science* *184*, 1192-1194.

Jonides, J. and H. Gleitman. (1972) A conceptual category effect in visual search: O as a letter or a digit. *Percept. Psychophys.* *12*, 457-460.

Kahneman, D. (1968) Method, findings, and theory in studies of visual masking. *Psychol. Bull.* *70*, 404-426.

Katz, J. J. (1971) *The Underlying Reality of Language and Its Philosophical Import*. (New York: Harper & Row).

Katz, L. and D. Wicklund. (1972) Word scanning rate in good and poor readers. *J. Educ. Psychol.* *62*, 138-140.

Kinsbourne, M. and E. K. Warrington. (1962a) The effect of an aftercoming random pattern on the perception of brief visual stimuli. *Quart. J. Exp. Psychol.* *14*, 223-234.

Kinsbourne, M. and E. K. Warrington. (1962b) Further studies on the masking of brief visual stimuli by a random pattern. *Quart. J. Exp. Psychol.* *14*, 235-245.

Kolers, P. A. (1964) Apparent movement of a Necker cube. *Amer. J. Psychol.* *77*, 220-230.

Kolers, P. A. (1966) An illusion that dissociates motion, object, and meaning. *Quarterly Progress Report* (Research Laboratory of Electronics, MIT) *82*, 221-223.

Kolers, P. A. (1967) Comments on the session on visual recognition. In *Models for the Perception of Speech and Visual Form*, ed. by W. Wathen-Dunn. (Cambridge, Mass.: MIT Press).

Kolers, P. A. (1968) Some psychological aspects of pattern recognition. In *Recognizing Patterns*, ed. by P. A. Kolers and M. Eden. (Cambridge, Mass.: MIT Press).

Kolers, P. A. (1970) Three stages of reading. In *Basic Studies in Reading*, ed. by H. Levin and J. Williams. (New York: Basic Books).

Kolers, P. A. and J. R. Pomerantz. (1971) Figural change in apparent motion. *J. Exp. Psychol.* *87*, 99-108.

Kroll, N. E. A., T. Parks, S. R. Parkinson, S. L. Bieber, and A. L. Johnson. (1970) Short-term memory while shadowing: Recall of visually and of aurally presented letters. *J. Exp. Psychol.* *85*, 220-224.

Lee, D. N. (1974) Visual information during locomotion. In *Perception: Essays in Honor of James J. Gibson*, ed. by R. B. MacLeod and H. L. Pick, Jr. (Ithaca, N. Y.: Cornell University Press).

Lee, D. N. and E. Aronson. (1974) Visual proprioceptive control of standing in human infants. *Percept. Psychophys.* *15*, 529-532.

Lishman, J. R. and D. N. Lee. (1973) The autonomy of visual kinaesthesis. *Perception* *2*, 287-294.

Mace, W. M. (1974) Ecologically stimulating cognitive psychology: Gibsonian perspectives. In *Cognition and the Symbolic Processes*, ed. by W. Weimer and D. S. Palermo. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Mace, W. M. (in press) James Gibson's strategy for perceiving: Ask not what's inside your head, but what your head's inside of. In *Perceiving, Acting and Comprehending: Toward an Ecological Psychology*, ed. by R. E. Shaw and J. Bransford. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

MacKay, D. M. (1967) Ways of looking at perception. In *Models for the Perception of Speech and Visual Form*, ed. by W. Wathen-Dunn. (Cambridge, Mass.: MIT Press).

Mackworth, A. K. (1973) Interpreting pictures of polyhedral scenes. In *Third International Joint Conference on Artificial Intelligence*. (Menlo Park, Calif.: Stanford Research Institute).

Maffei, L. and A. Fiorentini. (1972) Processes of synthesis in visual perception. Nature 240, 479-481.

Marcel, A. J. (1974) Perception with and without awareness. Paper presented at Experimental Psychology Society Meetings, Stirling, Scotland, July.

Marshall, J. D. and F. Newcombe. (1971) Patterns of paralexia. Paper presented at the International Neuropsychology Symposium, Engelberg, Switzerland.

Mayhew, J. E. W. and S. M. Anstis. (1972) Movement aftereffects contingent on color, intensity, and pattern. Percept. Psychophys. 12, 77-85.

McCulloch, W. S. (1945) A heterarchy of values determined by the topology of nervous nets. Bull. Math. Biophys. 1, 89-93.

McCullough, C. (1965) Color adaptation of edge-detectors in the human visual system. Science 149, 1115-1116.

Merleau-Ponty, M. (1947) Phenomenology of Perception, trans. by C. Smith. (New York: Humanities Press, 1962).

Minsky, M. (1963) Steps toward artificial intelligence. In Computers and Thought, ed. by A. E. Fergenbaum and J. Feldman. (New York: McGraw-Hill).

Minsky, M. and S. Papert. (1972) Artificial Intelligence, Memo 252. (Cambridge, Mass.: MIT Press).

Moray, N. (1967) Where is capacity limited? A survey and a model. Acta Psychol. 27, 84-92.

Murch, G. M. (1972) Binocular relationships in a size and color orientation aftereffect. J. Exp. Psychol. 93, 30-34.

Neisser, U. (1967) Cognitive Psychology. (New York: Appleton-Century-Crofts).

Novik, N. (1974) Developmental studies of backward visual masking. Unpublished Ph.D. thesis, University of Connecticut.

Ogle, K. N. (1962) Perception of distance and of size. In The Eye, vol. 4, ed. by H. Davson. (New York: Academic Press).

Pastore, N. (1971) Selective History of Theories of Visual Perception: 1650-1950. (New York: Oxford University Press).

Phillips, W. A. (1974) On the distinction between sensory storage and short-term visual memory. Percept. Psychophys. 16, 283-290.

Phillips, W. A. and A. D. Baddeley. (1971) Reaction time and short-term visual memory. Psychon. Sci. 22, 73-74.

Pittenger, J. B. and R. E. Shaw. (in press) Aging faces as viscal-elastic events: Implications for a theory of nonrigid shape perception. J. Exp. Psychol.: Human Perception and Performance.

Pollen, D. A., J. R. Lee, and J. H. Taylor. (1971) How does the striate cortex begin the reconstruction of the visual world? Science 173, 74-77.

Posner, M. I. (1966) Components of skilled performance. Science 152, 1712-1718.

Posner, M. I. (1969) Abstraction and the process of recognition. In Psychology of Learning and Motivation, vol. 3, ed. by G. H. Bower and J. T. Spence. (New York: Academic Press).

Posner, M. I., S. J. Boies, W. H. Eichelman, and R. L. Taylor. (1969) Retention of visual and name codes of single letters. J. Exp. Psychol. Monogr. 79, 1-17.

Posner, M. I. and S. W. Keele. (1967) Decay of visual information from a single letter. Science 158, 137-139.

Posner, M. I. and R. F. Mitchell. (1967) Chronometric analysis of classification. Psychol. Rev. 74, 392-409.

Reicher, G. M. (1969) Perceptual recognition as a function of meaningfulness of stimulus material. J. Exp. Psychol. 81, 275-280.

Riggs, L. A. (1973) Curvature as a feature of pattern vision. Science 181, 1070-1072.

Roberts, L. G. (1965) Machine perception of three-dimensional solids. In Electro-Optical Information Processing, ed. by J. T. Tippott. (Cambridge, Mass.: MIT Press).

Rock, I., F. Halper, and T. Clayton. (1972) The perception and recognition of complex figures. Cog. Psychol. 3, 655-673.

Ryle, G. (1949) The Concept of Mind. (London: Hutchinson).

Scharf, B. and L. A. Lefton. (1970) Backward and forward masking as a function of stimulus and task parameters. J. Exp. Psychol. 84, 331-338.

Schiff, W. (1965) Perception of impending collision: A study of visually directed avoidance behavior. Psychol. Monogr. 79, Whole No. 604.

Schiff, W., J. A. Caviness, and J. J. Gibson. (1962) Persistent fear responses in Rhesus monkeys to the optical stimulus of "looming." Science 136, 982-983.

Schiller, P. H. (1965) Monoptic and dichoptic visual masking by patterns and flashes. J. Exp. Psychol. 69, 193-199.

Sekuler, R. (1974) Spatial vision. In Annual Review of Psychology. (Palo Alto, Calif.: Annual Reviews, Inc.).

Selfridge, O. G. and U. Neisser. (1960) Pattern recognition by machine. Sci. Amer. 203 (Aug.), 60-68.

Shaw, R. E. (1971) Cognition, simulation, and the problem of complexity. J. Structural Learning 2, 31-44.

Shaw, R. E. and M. McIntyre. (1974) Algoristic foundations to cognitive psychology. In Cognition and the Symbolic Processes, ed. by W. Weimer and D. Palermo. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Shaw, R. E., M. McIntyre, and W. Mace. (1974) The role of symmetry in event perception. In Perception: Essays in Honor of James J. Gibson, ed. by R. B. MacLeod and H. L. Pick, Jr. (Ithaca, N. Y.: Cornell University Press).

Shaw, R. E. and J. Pittenger. (in press) Perceiving the face of change in changing faces: Toward an event perception theory of shape. In Perceiving, Acting, and Comprehending: Toward an Ecological Psychology, ed. by R. Shaw and J. Bransford. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Shaw, R. E. and B. E. Wilson. (in press) Generative conceptual knowledge: How we know what we know. In Cognition and Instruction: Tenth Annual Carnegie-Mellon Symposium on Information Processing, ed. by D. Klahr. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Shepard, R. N. (1962) The analysis of proximities: Multidimensional scaling with an unknown distance function. I and II. Psychometrika 27, 125-140 and 214-246.

Shepard, R. N. (in press) Studies of the form, formation, and transformation of internal representations. In Cognitive Mechanisms, ed. by E. Galanter. (Washington, D.C.: V. H. Winston & Sons).

Shepard, R. N. and S. Chipman. (1970) Second-order isomorphism of internal representation: Shapes of states. Cog. Psychol. 1, 1-17.

Shepard, R. N. and C. Feng. (1972) A chronometric study of mental paper folding. Cog. Psychol. 3, 228-243.

Shepard, R. N. and J. Metzler. (1971) Mental rotation of three-dimensional objects. Science 171, 701-703.

Sherrington, C. S. (1906) The Integrative Action of the Nervous System. (Cambridge, England: University of Cambridge Press).

Smith, M. C. and P. H. Schiller. (1966) Forward and backward masking: A comparison. Canad. J. Psychol. 20, 337-342.

Sperling, G. (1960) The information available in brief visual presentations Psychol. Monogr. 74, Whole No. 498.

Sperling, G. (1963) A model for visual memory tasks. Human Factors 5, 19-31.

Sperling, G. (1967) Successive approximations to a model for short-term memory. Acta Psychol. 27, 285-292.

Sternberg, S. (1966) High-speed scanning in human memory. Science 153, 652-654.

Sternberg, S. (1967) Two operations in character-recognition: Some evidence from reaction-time measurements. Percept. Psychophys. 2, 45-53.

Sternberg, S. (1969) Memory scanning: Mental processes revealed by reaction time experiments. Amer. Scient. 57, 421-457.

Stromeyer, C. F. and R. J. Mansfield. (1970) Colored aftereffects produced with moving images. Percept. Psychophys. 7, 108-114.

Sutherland, N. S. (1973) Intelligent picture processing. Paper presented at the Conference on the Evolution of the Nervous System and Behavior, Florida State University, Tallahassee.

Thompson, D. W. (1917) On Growth and Form, 2nd ed. (Cambridge, England: University of Cambridge Press, 1942).

Treisman, A., R. Russell, and J. Green. (1974) Brief visual storage of shape and movement. In Attention and Performance, vol. 5. (New York: Academic Press).

Turvey, M. T. (1972) Some aspects of selective readout from iconic storage. Haskins Laboratories Status Report on Speech Research SR-29/30, 1-14.

Turvey, M. T. (1973) On peripheral and central processes in vision: Inferences from an information processing analysis of masking with patterned stimuli. Psychol. Rev. 80, 1-52.

Turvey, M. T. (1974) Constructive theory, perceptual systems, and tacit knowledge. In Cognition and the Symoblic Processes, ed. by W. Weimar and D. Palermo. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Turvey, M. T. and S. Kravetz. (1970) Retrieval from iconic memory with shape as the selection criterion. Percept. Psychophys. 8, 171-172.

Vanthoor, F. L. J. and E. G. J. Eijkman. (1973) Time course of the iconic memory signal. Acta Psychol. 37, 79-85.

von Fieandt, K. and J. J. Gibson. (1959) The sensitivity of the eye to two kinds of continuous transformation of a shadow-pattern. J. Exp. Psychol. 57, 344-347.

von Wright, J. M. (1968) Selection in immediate memory. Quart. J. Exp. Psychol. 20, 62-68.

von Wright, J. M. (1970) On selection in visual immediate memory. Acta Psychol. 33, 280-292.

Walker, J. T. (1972) A texture-contingent visual motion aftereffect. Psychon. Sci. 28, 333-335.

Wallach, H. and D. N. O'Connell. (1953) The kinetic depth effect. J. Exp. Psychol. 45, 205-207.

Warren, R. (1975) The perception of egomotion. Unpublished Ph.D. thesis, Cornell University.

Weisstein, N. (1969) What the frog's eye tells the human brain: Single cell analyzers in the human visual system. Psychol. Bull. 72, 157-176.

Weisstein, N. (1970) Neural symbolic activity: A psychophysical measure. Science 168, 1489-1499.

Weisstein, N. A. (1971) W-shaped and U-shaped functions obtained for monoptic and dichoptic disk-disk masking. Percept. Psychophys. 9, 275-278.

Weisstein, N. and C. S. Harris. (1974) Visual detection of line segments: An object-superiority effect. Science 186, 752-754.

Weisstein, N., F. Montalvo, and G. Ozog. (1972) Differential adaptation to gratings blocked by cubes and gratings blocked by hexagons: A test of the neural symbolic activity hypotheses. Psychon. Sci. 27, 89-92.

Weyl, H. (1952) Symmetry. (Princeton, N. J.: Princeton University Press).

Wheeler, D. D. (1970) Processes in word recognition. Cog. Psychol. 1, 59-85.

White, B. W., F. A. Saunders, L. Scadden, P. Bach-Y-Rita, and C. C. Collins. (1970) Seeing with the skin. Percept. Psychophys. 7, 23-27.

Wickens, D. D. (1972) Characteristics of word encoding. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington, D.C.: V. H. Winston & Sons).

Winograd, T. (1972) Understanding natural language. Cog. Psychol. 3, 1-191.

The Perception of Speech*

C. J. Darwin+

ABSTRACT

The general plan of this paper is first to examine what problems our ability to perceive speech raise and then to look at some of the mechanisms proposed as possible solutions to these problems. The section on the structure of speech outlines the articulatory foundation for the basic invariance problems in speech perception; the next section on cues to phonemic categories illustrates the perceptual correlates of these articulatory constraints; while the section on auditory grouping and feature extraction describes some of the mechanisms that may be operating at early stages in auditory and speech perception. The last section looks briefly at suprasegmental factors, such as prosody, and weighs the evidence for various models of speech perception.

INTRODUCTION

One of the most striking phenomena in the perception of speech is the degree to which our conscious experience follows the semantic intention of the speaker. Our conscious perceptual world is composed of greetings, warnings, questions, and statements; while their vehicle, the segments of speech, goes largely unnoticed and words are subordinated to the framework of the phrase or sentence. Nor is this "striving after meaning" a mere artifact, confined to situations in which we want to understand rather than to analyze, since our ability to analyze speech into its components is itself influenced by higher-level units. The basic physical dimensions of the stimulus and even, under normal listening conditions, the segments of speech are inaccessible to us directly. This is evident in experiments on the perception of prosody and in

---

*To appear in Handbook of Perception, vol. 7, ed. by E. C. Carterette and M. P. Friedman. (New York: Academic Press).

+Also University of Connecticut, Storrs; on leave of absence from University of Sussex, Brighton, England.

Acknowledgment: Part of this chapter was written while the author was on leave of absence at the University of Connecticut and Haskins Laboratories. He is indebted to Alvin Liberman, Gary Kuhn, Ignatius Mattingly, Peter Eimas, and Susan Brady for discussion and comments, to Michael Studdert-Kennedy for putting him on the trail of the Troyan strumpet, and to his wife Kate for finding her for him.

experiments in which subjects listen for a particular phonemic segment. The results of these experiments parallel in an interesting way the perception of three-dimensional objects in vision.

Lieberman (1965), for example, studied whether the intonation contours transcribed by linguists trained in a particular transcription system corresponded with the objective pitch present in the stimulus. He found that many of the pitch features described by his linguists were not in fact present but were introduced by the perceived syntactic structure of the sentence. The pitch the linguists heard was often that which their system told them should accompany a particular syntactic structure. Somewhat similar results have been reported by Hadding-Koch and Studdert-Kennedy (1964). They found that subjects' judgments about whether a short utterance has a rise in pitch at the end or not were influenced by the pitch contour in the earlier part of the utterance in a way that is thought to reflect aerodynamic constraints on spoken pitch (Lieberman, 1967). In a similar vein, Huggins (1972a) found that subjects had lower thresholds for detecting changes in the duration of segments of speech from "normal" when they based their judgments on the rhythm or stress of the utterance than if they tried to listen for duration itself. Even such familiar dimensions as pitch and duration seem to be relatively inaccessible to conscious experience. There is perhaps a parallel here with the problems that face a painter trying to represent a three-dimensional object in two dimensions. Gombrich (1960) describes how different stylistic schools resort to a repertoire of visual tricks or schemata to achieve the two-dimensional representation of a scene, attaining, like Lieberman's linguists, a representation determined as much by their training as by the object portrayed. In both cases a more objective representation can be attained by resorting to artificial aids, which distract attention from the meaning of the elements portrayed. Dürer illustrates one extreme way of doing this in his woodcut of a draftsman drawing a reclining nude by viewing her from a fixed point in space through a transparent grid that corresponds to a similar grid on his canvas (Gombrich, 1960:306). In the same way Daniel Jones transcribed intonation contours by lifting the needle from a phonograph record in midsentence and judging the last-heard pitch. In both these cases some external segmentation has been imposed on the perceptual whole to remove some of the changes caused by semantic interpretation. Without this aid the listener can only reconstruct properties of the stimulus from what he knows about higher-level categories.

This reconstruction also appears to be operating in experiments in which the subject is asked to listen for a particular speech segment. Following traditional phonetics, we will use the term "phoneme" to refer to the smallest segment of speech that distinguishes two words of different meaning. The intuitive value of this concept outweighs, for the purposes of this paper, the theoretical problem that it raises. The words "kine" and "Cain" thus share initial and final phonemes (/k/ and /n/, respectively) but differ in the medial diphthong. When subjects listening to a list of two-syllable words are asked to press a key whenever they hear a particular phoneme, they do this more slowly than when they are asked to detect a whole syllable; this in turn is slower than detecting a whole word (Savin and Bever, 1970; Foss and Swinney, 1973). Moreover, the time to detect a given phoneme at the beginning of a syllable is shorter if that syllable is a word than if it is not (Rubin, Turvey, and Van Gelder, in press). Here the subject's performance is influenced by larger units of analysis despite the needs of the task. If these higher levels are removed so that phonemes are being targeted for in lists of phonemes, and syllables in lists of syllables, then the smaller units are in fact detected faster than the larger (McNeill and Lindig, 1973). Our conscious awareness, then, is driven to the highest level present in the stimulus, allowing lower levels to be accessible only as a subsequent operation on these higher units.

This somewhat paradoxical finding, that conscious decisions at a higher level can be made before decisions at a lower level, though striking in speech, is also encountered in written language. Letters embedded in words are identified more accurately than letters in nonword strings or in isolation (Reicher, 1969; Wheeler, 1970) provided, again paradoxically, that subjects do not know in which position in the word the letter that they have to report will occur (Johnston and McClelland, 1974). Although both the speech and the written language results demand an interpretation that stresses the importance of higher-order units in perception, there are interesting differences between the two modalities.

First, subjects are quicker at detecting an isolated phoneme than one embedded in a word (McNeill and Lindig, 1973); in vision isolated letters are identified less accurately than those in words (Wheeler, 1970). Second, knowing where in a word a target will appear does not abolish the relative difficulty of detecting that phoneme, compared with detecting the whole word (Foss and Swinney, 1973; McNeill and Lindig, 1973); in vision, on the other hand, knowing where in a string the target letter will appear reverses the usual advantage of the string being a word (Johnston and McClelland, 1974). This latter discrepancy may perhaps be taken as illustrating the relative ease with which spatially defined portions of a visual stimulus may be attended to compared with temporally defined portions of an auditory stimulus. A more attractive explanation is that these differences arise because of the basically different nature of spoken and written language. Embedding a phoneme in a syllable, or a syllable in a word, is, as we shall see, very different from embedding a letter or string of letters in a type-written word. The whole physical representation of the phonemic element can be completely restructured in a way that prevents the subject from being able to attend to it or detect it without also taking account of the context in which it occurs.

To appreciate the implications of this restructuring for perception, let us follow the changes that take place in semantic elements as they are described linguistically at different levels. The choice of linguistic systems is to some extent arbitrary, and we will dip both into contemporary generative phonology and into acoustic and articulatory phonetics.

THE STRUCTURE OF SPEECH

There is growing psychological evidence that the morpheme is a significant unit in word recognition. It is more powerful than the word at explaining frequency effects in perception (Murrell and Morton, 1974), and seems to be able to explain effects of word length on perception that have previously been attributed to number of syllables.[1] This evidence is primarily from studies of the written word but it is probable that similar effects could also be obtained for speech, particularly as in generative phonology the systematic phonemic level bears striking similarities to the written word. The systematic phonemic level represents the linguistic message as an ordered sequence of elements that preserve morphemic invariance, but that can be mapped, using the particular rules of a language or dialect, into a phonetic representation. A speech synthesis-by-rule program could then be used to derive an intelligible utterance from such a representation. At the systematic phonemic level the words "courage" and "courageous" can be represented as /koræge/ and /koræge+as/, forms that bear an interesting resemblance to the spelled word (Chomsky and Halle, 1968; Chomsky, 1970; Gough, 1972). Quite specific to speech though is the restructuring that takes place between this systematic phonemic level and the phonetic level to give strings such as [kʌ́rəj] and [kəre̅́y̅ǰəs] (Chomsky and Halle, 1968:235), forms from which an articulatory specification and thence an acoustic signal could be derived. This variation between the systematic phonemic level and the phonetic is one that, though not generally accessible to the naive listener is nonetheless readily distinguished by a phonetician. In the example cited above the change in the second vowel is not difficult to apprehend, but more subtle changes are covered by the same level of rules (such as the change in aspiration in the /p/ in "pit" and "spit"), which are difficult to perceive as such except by the trained ear. This variation has been termed extrinsic allophonic variation (Wang and Fillmore, 1961; Ladefoged, 1966) as distinct from intrinsic, which is not accessible even to the trained ear of the phonetician. To appreciate the problems raised by intrinsic allophonic variation we need to consider in outline the mechanisms by which speech is produced.

---

[1] Max Coltheart and Roger Freeman: personal communication.

In normal (rather than whispered) speech, the main source of sound is the vibration of the vocal cords. These are set into vibration by airflow from the lungs, the frequency of the vibration being determined jointly by the stiffness of the cords and the pressure drop across them (see MacNeilage and Ladefoged, in press). The waveform of the sound at this stage is roughly an asymmetrical triangle whose spectrum, for a continuously held pitch, is a series of harmonics of the fundamental, decreasing in amplitude with increasing frequency. The effect of the cavities of the mouth (and of the nose when they are coupled in, as during nasal consonants and nasalized vowels) is to change this spectrum so that well-defined broad peaks occur in it, corresponding to the resonant frequencies of the system of cavities. These broad peaks are formants and their frequencies and amplitudes vary in a well-understood way (Fant, 1960) with the changing shape of the vocal tract, as the various articulators move to give formant transitions. The values of the formants are independent of the pitch of the voice, although the accuracy with which a formant peak can be estimated from the harmonic structure depends on the particular pitch present. For unvoiced consonants such as [p,f], or in whispered speech, the vocal cords do not vibrate but instead the sound source is noise from turbulent air either at the glottis (for [h] aspiration and whisper) or at some other point of constriction. Peaks corresponding to the vocal-tract resonant frequencies still exist, but the spectrum will be continuous rather than composed of harmonic lines. The peaks found in the noise spectrum will reflect mainly the resonances of cavities in front of the point of constriction at which the noise is generated.

The acoustic speech signal reflects only indirectly the movements of the individual articulators of the vocal tract. These articulators can move with a large degree of independence: the lips move independently of the tongue, different parts of the tongue have some independence (Öhman, 1967), and whether the vocal cords vibrate or not depends but little on the movements of the supralaryngeal tract provided the airflow is not blocked. This ability of independent movement of the articulators would present little problem if speech consisted of a sequence of rigidly defined vocal-tract positions, with every distinctive linguistic unit having a concomitant vocal-tract configuration. Unfortunately, this is not the case. Only in the production of continuously held vowels does the vocal tract come close to this ideal, in that a change in the position of any articulator will modify the quality of the vowel. For consonants, though, the situation is much more complicated. They are produced by constricting the vocal tract at a particular place--the place of articulation--and in a particular manner. The constriction formed by different manners of articulation can either be complete (as for the stops [b,d,g,p,t,k]), complete for the oral cavities but with the soft palate lowered to allow airflow through the nose (as for the nasals [m,n,ŋ]), incomplete but sufficiently close to give turbulent noise (as in fricatives [f,s,ʃ,v,z,ʒ]), complete closure in the midline but with space for airflow at the sides (as with the lateral [l]), or so incomplete as to approximate extreme vowel positions (as in the semivowels /w,j/). Definition of the place and manner of the articulation is sufficient together with voicing to define the consonant; provided this articulation is accomplished, the articulatory mechanisms not involved with this gesture are free to get on with whatever the forthcoming phonemes require. It is this property of coarticulation that is one source of the intrinsic variation between the signal and the phoneme. A well-known extreme example is that in the syllable /stru/ the lips are free to round in anticipation of the vowel *three* phonemes later irrespective of word boundaries (Daniloff and Moll, 1968). Similar and perceptually significant coarticulations occur with nasality (Ali, Gallagher, Goldstein, and Daniloff, 1971; Moll and Daniloff, 1971), which are also unconstrained by word boundaries (Dixit and MacNeilage, 1972).

A related source of variation is that for most of the time the vocal tract is moving from one target position to another so that information about which targets are being approached or left is carried by transitional information, which of its very nature depends both on the immediately adjoining and more distant targets. For example, the formant transitions produced by the rapid movement away from a point of constriction into a subsequent vowel can be a sufficient cue to the location of that point of constriction (Cooper, Delattre, Liberman, Borst, and Gerstman, 1952).

A further complication comes from the articulation of vowels in normal fast speech being considerably less differentiated than in the citation of individual words (Shearme and Holmes, 1962). In addition, in rapid speech, the target articulation position for a vowel may not be reached before the tongue must move back toward the target position for the next phoneme. This articulatory undershoot necessarily implies a corresponding acoustic undershoot so that instead of reaching and holding a target position the formants go through maxima and minima short of their target (Lindblom, 1963).

These examples illustrate some of the production mechanisms responsible for "intrinsic" allophonic variation. But they do not exhaust the sources of variation; for, as Studdert-Kennedy (1974) points out, between-speaker but within-dialect variation is not covered by either of the categories intrinsic or extrinsic. The problem here is that different speakers have different-sized heads, so that the formants produced when a child articulates a particular vowel are higher in frequency than those produced by an adult. Even within normal adults there is a variation of about 20 percent (Peterson and Barney, 1952). Moreover, this is not a simple scaling problem. The values of the individual formants for different vowels change by different proportions for male and female speakers in a way that suggests that men have proportionately larger pharynxes than women (Fant, 1966).

CUES TO PHONEMIC CATEGORIES

There is abundant evidence that articulatory mechanisms structure the acoustic signal in such a way that in general there is no simple relationship between a phonetic category and those sounds that are sufficient to cue it. Work on the perception of synthetic speech has made this point elegantly, enabling us to understand the relationship between phonetic categories and their cues sufficiently well to produce intelligible speech automatically from a phonetic symbol input (Liberman, Ingemann, Lisker, Delattre, and Cooper, 1959; Holmes, Mattingly, and Shearme, 1964; Kuhn, 1973). What is perhaps less well understood is the degree to which the cues that have been shown to be important in synthetic speech are also important in natural speech. Subjects can be variable in the way they react to synthetic speech and yet natural speech retains its intelligibility under a wide range of distortions. It is possible that this discrepancy is due to natural speech containing a wider variety of cues than has been investigated systematically with synthetic speech and also to different listeners using the particular cues that are used in the synthesis to different extents. But another factor is undoubtedly that we do not yet understand sufficiently the changes in the cues to segments that occur with context. While this has been emphasized for such effects as that of the neighboring vowels on a consonant (Delattre, Liberman, and Cooper, 1955; Öhman, 1966), other interactions--such as cues to consonants in clusters, changes that occur with word boundaries, and changes with stress patterns and speaking rate--have received very little attention in perceptual experiments until quite recently.

Vowels

Although vowels produced in isolated syllables can be adequately distinguished by the steady-state values of their first two or three formants, it is unlikely that vowel perception in running speech can be dealt with in such a simple way. Normal continuous speech introduces at least two complications: speaker change and rapid articulation. Using synthetic two-formant patterns, Ladefoged and Broadbent (1957) showed that varying the range of frequencies used in a precursor sentence influenced the vowel quality attributed to a fixed formant pattern at the end of the sentence. Similar effects have been found for consonants (Fourcin, 1968). A particular formant pattern is thus perceived relative to some frame that is characteristic of the particular speaker. Information about this frame can be provided either by a precursor sentence or, it seems, by the syllable that contains the test vowel itself. Shankweiler, Strange, and Verbrugge (in press) have shown that a potpourri of vowels produced by many different speakers is more intelligible if the vowels are flanked by consonants than if they are spoken in isolation without the consonants. The dynamic information from the consonant transitions may limit the possible tract configurations that could have produced the syllable more effectively than a steady state could, or it may simply allow the formants to be detected more accurately.

Embedding a vowel between consonants can, at rapid rates of articulation, prevent the articulators and hence the formant pattern from reaching the target values (Lindblom, 1963). Nevertheless, the perceptual system appears able to compensate for this, so that a vowel is perceived whose steady-state formant values would have been more extreme than those actually reached in the syllable (Lindblom and Studdert-Kennedy, 1967). This is illustrated in Figure 1.

Stops

Stop consonants in intervocalic position are characterized by an abrupt and complete closure (or stopping) of the vocal tract at some point of articulation, followed by an abrupt release of this closure. With voiceless stops the vocal folds cease vibrating at closure and do not start again until some time after release. The period of closure is thus silent. For the voiced stops though, the vocal cords continue to vibrate throughout the period of closure (provided it is not too long) producing a low-amplitude, low-frequency sound. The abrupt changes in amplitude at closure and release are cues to the presence of a stop consonant; /slit/ will change to /split/ if a brief period of silence is introduced between the friction and the vowel (Bastian, Eimas, and Liberman, 1961). Other distinctive acoustic events indicating the presence of a stop consonant are the burst of energy at release and, for voiced stops, the rapid rise in the first formant after release.

When a stop is released, there is a sudden drop in pressure in the mouth cavity and a rapid flow of air through the widening constriction. The initial drop in pressure gives an impulsive excitation to the mouth cavities that is followed by a brief period of frication from the turbulent airflow through the constriction. As the constriction widens, the airflow becomes smooth and then the only source of excitation is that from the glottis, which for voiced sounds will be air pulses, and for voiceless sounds noise produced by turbulent airflow at the glottal constriction. The spectrum of the emergent sound is predominantly determined by the cavities in front of the source of energy. Formant structure similar to that for voiced sounds appears in the noise originating from the glottis (aspiration), but to a much lesser extent in the noise originating at a supraglottal place of articulation (burst and frication), which reflects predominantly the resonant frequency of the cavity in front of the place of articulation. The formant pattern changes as the articulators move away from the point of articulation into the position appropriate for the next segment.

A voiceless stop can be cued simply by putting a burst of noise at a suitable frequency, a short time in front of a vowel. What place of articulation is heard depends on the frequency of the noise and on the vowel that follows. It is possible to make the same burst of noise sound like two different consonants by placing it before different vowels [e.g., /pi/, /ka/, /pu/ (Liberman, Delattre, and Cooper, 1952)]. Similarly the significance in the change of the formant pattern depends crucially on what vowel the formants lead to (Delattre, Liberman, and Cooper, 1955). The reason for this dependency is partly that the formants have to lead into the vowel, but it is also because during closure the articulators are in a position that anticipates the forthcoming vowel. Because of this coarticulation, the formant pattern at release will vary with the vowel. A recent paper by Kuhn (1975) points out that the formant transition that appears to carry the burden of cuing the place of articulation of the stop is the one that for a particular stop-vowel combination is associated with the mouth cavity. The curiosity of the burst of noise at 1400 Hz that cues /pi/,

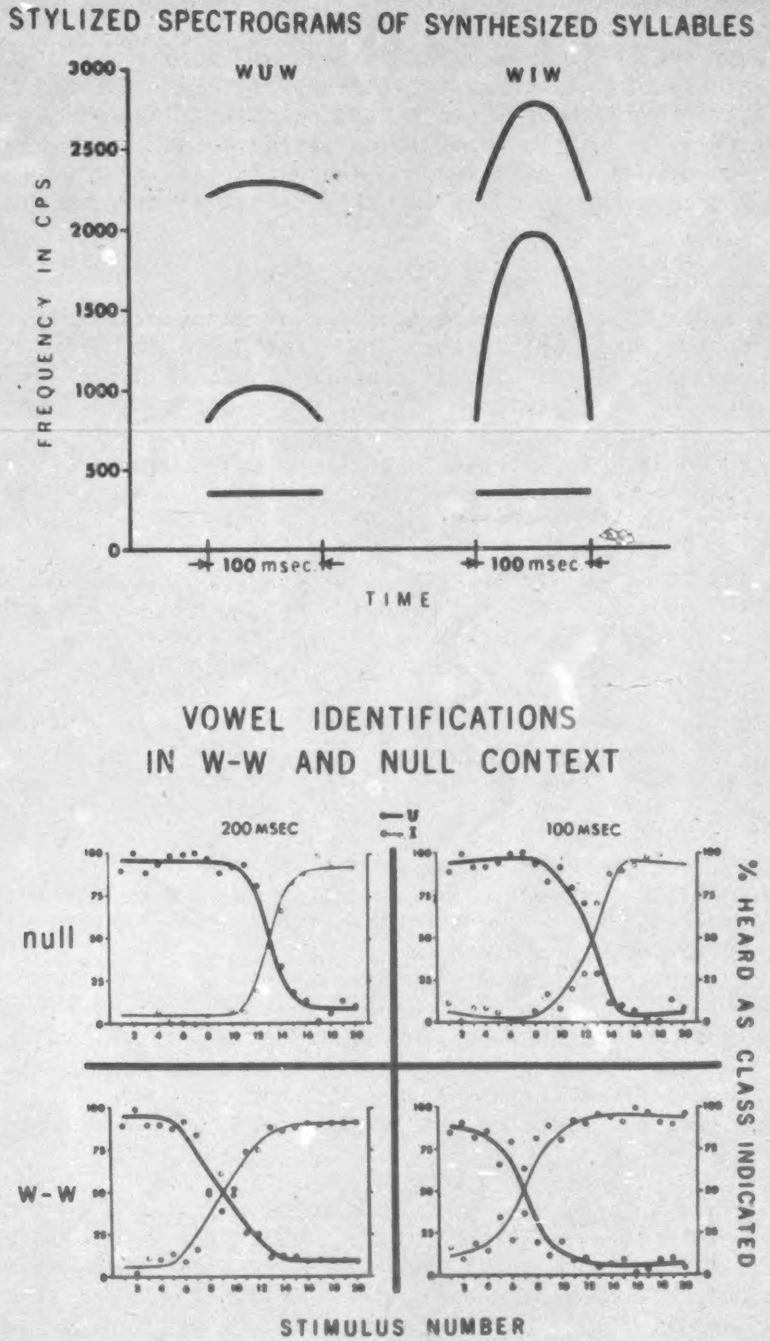

Figure 1: The upper panel shows stylized spectrograms of synthetic syllables differeing in the maximum values attained by the second and third formants. The lower panel shows how subjects heard vowels distributed along a continuum between these maxima. The null condition vowels lack transitions: they have steady-state formants at the maximum values attained by the "w-w" patterns. As may be seen from the leftward shift of the /ʊ-ɪ/ boundary in the "w-w" series in the same way as steady-state patterns having formant values beyond their maxima. This "overshoot" is greater for fast (100 msec) than for slow (200 msec) patterns (Figures 2 and 3, Studdert-Kennedy and Cooper, 1966; reproduced with permission of publisher and authors).

/ka/, and /pu/ is described by Kuhn as follows: "Before /i/ this burst appears to be interpreted as part of the rise in frequency of the front cavity resonance as it moves up to F_3. Before /a/, the burst appears to be interpreted as part of the fall in frequency of the front cavity as it moves to a slightly lower value in F_2. Before /u/ it appears to be interpreted as part of a flat, lip-release spectrum and was a somewhat worse cue."

For intervocalic stops with different vowels on either side (e.g., /idu/) it has been shown by Öhman (1966) that the formant transitions both into and out of the stop closure are jointly influenced by both vowels. However, the perceptual significance of this observation is not entirely clear. Two independent experiments have failed to find any perceptual correlate of this coarticulation effect. Fant, Liljencrants, Malac, and Borovickova (1970) using both natural and synthetic speech could find no general effect of this coarticulation on intelligibility, and Lehiste and Shockey (1972) found that listeners could not judge the missing vowel in the vowel-consonant (VC) part of natural vowel-consonant-vowel (VCV) utterances, even though they claim that coarticulation effects similar to those observed by Öhman (1966) could be seen in spectrograms of their stimuli. These two experiments are particularly interesting since they are a rare example of perception seemingly not being sensitive to articulatory constraints.

Invariant Cues for Stops?

The issue has been revived recently by Cole and Scott (1974a, 1974b) whether in natural speech there exist invariant cues or combinations of cues uniquely specifying particular consonants independent of the succeeding vowel. The claim made by Cole and Scott is that "place of articulation [for stops] is signaled by a set of cues which form invariant patterns for /b/, /d/, /p/, /t/, /k/ in initial position before a vowel in a stressed syllable..." (1974b:359).

This claim is clearly at odds with the results from synthetic speech showing that neither the burst (Liberman, Delattre, and Cooper, 1952) nor the formant transitions (Delattre, Liberman, and Cooper, 1955) of adequate synthetic syllables are invariant with vowel context. One resolution of this difference is to presume that the synthetic speech studies have missed some important cue or combination of cues that are responsible for invariant perception in natural speech. A closer inspection of the data on which Cole and Scott's conclusion is based, however, suggests that there is no real discrepancy. Their conclusion is made possible only by their being selective in the results they consider, by a loose interpretation of "invariant," and by an ambiguous use of the term "burst."

Briefly the evidence is this: if the burst (excluding the aspiration) from a natural voiceless stop produced in syllable-initial position before one vowel is spliced onto a different vowel containing a brief period of aspiration without formant transitions (as formed in producing /h/), then a stop is heard whose place of articulation will vary with the choice of vowels in a way that is consistent with the results from synthetic speech (Liberman, Delattre, and Cooper, 1952; Schatz, 1954). In particular, the same burst will be heard as /p/ before /i/ and /u/, but as /k/ before /a/. If instead of just the burst a longer portion of the sound from a voiceless consonant is removed and translated onto a different steady-state vowel, then again subjects' percepts do not change if the vowels interchanged are /i/ and /u/, but they are also relatively little affected if the vowel /a/ is included (Cole and Scott, 1974a). The discrepancy here is readily explicable if we examine what information is being translated from one vowel context to another. In Cole and Scott's experiment the duration of sounds translated were: /b/ 20 msec, /d/ 30 msec, /g/ 40 msec, /p/ 50 msec, /t/ 80 msec, and /k/ 100 msec. With the possible exception of /b/ and /d/, these durations are sufficient to give considerable information about the following vowel by virtue of formant transitions within the aspiration. Indeed, an experiment cited by Cole and Scott (1974a) in support of their claim demonstrates this point. Winitz, Scheib, and Reeds (1972) removed all but the burst and aspiration from natural tokens of /p,t,k/ spoken before the vowels /i,a,u/, and played the resulting sounds to subjects for identification either in isolation or followed by a 100-msec steady state of the *same* vowel before which they had been spoken. The mean durations of the translated segments were 70, 77, and 93 msec for /p,t,k/, respectively. Subjects were asked to identify in both these types of sound either the consonant or the vowel. The results showed that for the sounds consisting only of the burst and aspiration the correct consonant was identified 65 percent of the time and the correct *vowel* 64 percent. Adding the steady-state vowel raised the scores to 71 and 86 percent, respectively. In these data the consonant is no more identifiable than the vowel. Cole and Scott's (1974a) procedure is sufficiently close to that of Winitz et al. to let us presume that similar results would have been obtained with Cole and Scott's sounds, had they asked their subjects to identify the vowel in the isolated sounds. Although Cole and Scott claim that only the burst was translated, the durations used in their sounds clearly allows the acknowledged possibility that their sounds contained formant transitions. However, they claim (*pace* Winitz, Scheib, and Reeds) that these transitions should not have aided perception, citing the claim by Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967: 436) that formant transitions are not commutable between vowels. In fact, the claim made by Liberman and colleagues is that in formant transition patterns there is no commutable *stop-consonant segment*, since a slice from the beginning of a formant pattern of a complete syllable will either be heard as a nonspeech sound, or as a stop consonant *followed by some vowel*. This does not of course imply that following formant transitions by an inappropriate steady-state vowel prohibits the perceptual system from using the transition information to interpret the previous burst, just as it does when there is no additional vowel spliced on. Indeed, for some of the stimuli used by Cole and Scott (1974a) it is likely that listeners heard *two* vowels.

The apparent discrepancy between Cole and Scott's experiment, on the one hand, and those of Schatz (1954) and Liberman et al. (1967) on the other, is thus attributable to the longer stimuli, excised by Cole and Scott, carrying information about the following vowel. Since these stimuli probably contained sufficient information for the following vowel to be identified, Cole and Scott are no closer to demonstrating perceptual invariance of stop consonants with vowel context than if they had removed none of the vowel.

Fricatives, Nasals, and Liquids

The perception of fricatives and nasals is less controversial. In both fricatives and nasals in syllable-initial position there is a long period of steady state followed by transitions into the following vowel. The nature of the steady state provides the main cue to the manner of articulation of the consonant and also provides some information on the place of articulation. The nasal murmurs produced at different places of articulation are quite similar since the oral cavity acts only as a side chamber to the nasal cavities from which the sound is radiated. The different nasals *can* be distinguished on the basis of their nasal murmurs alone, but the bulk of the place information is carried by the formant transitions (Liberman, Delattre, Cooper, and Gerstman, 1954; Malecot, 1956).

In contrast to nasals, fricative spectra (with the exception of /f,θ/ and their voiced cognates /v,ð/) are markedly dissimilar (Strevens, 1960) and carry the bulk of the perceptual load for place of articulation. Harris (1958) segmented naturally spoken fricative-vowel syllables into a fricative portion and a transition + vowel portion, which were then commuted. She found that except for the /f,θ/ and /v,ð/ distinctions the place of articulation was perceived according to the fricative spectrum rather than the formant transitions.

Steady-state friction is the nearest that speech comes to a one-to-one mapping between sound and phonetic category. Yet even here there is variation with speaker and context, which, though not sufficient to cause confusion between the fricative categories, can serve to distinguish the sex of the speaker (Schwartz, 1968) and, through weak-formant transitions within the noise, to cue place of articulation of subsequent stops (Schwartz, 1967).

The liquids /r,l/ are characterized by having a brief (or in some contexts nonexistent) steady state with a low first formant, followed by a rather slow transition into the following vowel. The speed of the transition together with changes in the second and third formant cue the presence of the liquid segment, but /r/ seems to be distinguished from /l/ primarily by changes in the third formant (Lisker, 1957b; O'Connor, Gerstman, Liberman, Delattre, and Cooper, 1957).

Voicing

The dimension of voicing, which has received intensive study recently, provides perhaps the best example of the intimate relationship between the articulatory-acoustic constraints that shape the stimulus and the mechanisms used to perceive the phonemic category. In final consonants voicing can be cued by the duration of the preceding vowel (Denes, 1955; Raphael, 1972) and in poststressed intervocalic stops by the duration of the stop closure (Lisker, 1957a), however, the perception of stops in utterance initial prestressed position has received the most attention.

Lisker and Abramson (1964) exploited the concept of voice onset time (VOT) to describe the differences in speech production between the various categories of voicing in stop consonants. This dimension classifies a particular stop utterance according to the time difference between the vocal cords starting to vibrate and the stop closure being released. For English voiced stops [b,d,g] in utterance initial position, this time is usually either around zero or negative (the vocal folds starting to vibrate before the stop is released), while for the English voiceless aspirated stops [p^h,t^h,k^h] (as in pot, tot, cot) there is a lag of between 20 and 100 msec from the release of the stop to the onset of voicing. The value of this dimension is twofold. First it adequately describes the categorization of differently voiced stops from many languages, different stops from similar context falling into clusters along the VOT continuum, which are nonoverlapping for stops at the beginning of isolated words and show only slight overlap for words spoken in sentences (Lisker and Abramson, 1967). Second it explains the changes in the acoustic cues accompanying different stop-consonant voicings. A voiceless stop will normally have a more intense burst, a longer period of aspiration, and a weaker first formant during the aspiration than the voiced homolog. All these changes can be adequately explained by the effect of changing the time for which the vocal folds are held apart, inhibiting voicing. Abduction of the vocal folds allows a greater pressure to build up within the oral cavity, leading to a stronger burst on release. In addition, it provides only noise excitation, which is weaker in low-frequency energy than in voicing, and it acoustically couples the trachea to the oral cavities; these last two factors are responsible for reducing the intensity of the first formant. Spectrograms of real speech illustrating three different values of VOT are shown in Figure 2(a).

Perceptually, this complex of cues resulting from a single articulatory dimension raises the question of which cues are used and how are they combined. Lisker and Abramson (1970) have shown that synthetic syllables differing in VOT can be perceived as differing in voicing. But the syllables they used varied a number of the concomitant acoustic cues including both the actual time of onset of voiced excitation and the intensity of the first-formant transition. It can be shown that both these acoustic correlates of voice onset time are perceptually important. It was clear from the early synthetic studies of voicing that to produce a good token of /p,t,k/ there had to be both aspiration present during the initial period of formant transitions and a reduction (or cutback) in the amplitude of the first-formant transition (Liberman, Delattre, and Cooper, 1958). The importance of the first-formant transition has again been stressed in recent work by Stevens and Klatt (1974) and by Summerfield and Haggard (1974). At a given VOT, perceived voicing can be influenced by a first-formant transition after the onset of voicing or by the amount of energy in the aspirated portion of the first formant. Interesting questions are raised by the known variation in VOT with such contextual factors as rate of articulation (Lisker and Abramson, 1967; Summerfield, 1974) and the nature of the following segment (Lisker, 1961; Klatt, 1973). Although the use of the first-formant transition as a cue may allow the increase in VOT brought about by the presence of a liquid after the stop to be compensated for directly (Darwin and Brady, 1975), it is also likely that other context effects demand a change in the weightings attached to the various cues (Summerfield and Haggard, 1974).

Syllable Boundaries

A phonemic representation of speech needs to include some way of representing the perceptible difference between such phrases as "I scream" and "ice-cream." This is done in both traditional and generative phonology by introducing a juncture marker /#/ (/aɪ#skrim/ vs. /aɪs#krim/). The phonetic changes accompanying a change in juncture have been studied by Lehiste (1960) and by Gårding (1967), who show that some changes in juncture are perceptually easy to distinguish, but they do not show directly which cues are actually used. One of the most obvious allophonic changes that occurs with a change in juncture is that of the aspiration of voiceless stop consonants. When Helen enters in Act IV, scene 5, of *Troilus and Cressida*, she is greeted by a fanfare and the shout "The Troyans' trumpet!" The aspiration of the word-initial /t/ and the voicing of the word-final /s/ are the phonetic cues that potentially protect her from insult. Using politically less sensitive material, Christie (1974) has shown that the aspiration of voiceless stops *is* used as a cue to juncture, since adding aspiration to a /t/ in the context /asta/ increases by about 30 percent listeners' judgments that the syllable boundary occurs after, rather than before, the /s/. Malmberg (1955) has also shown that the presence of formant transitions

FIGURE 2

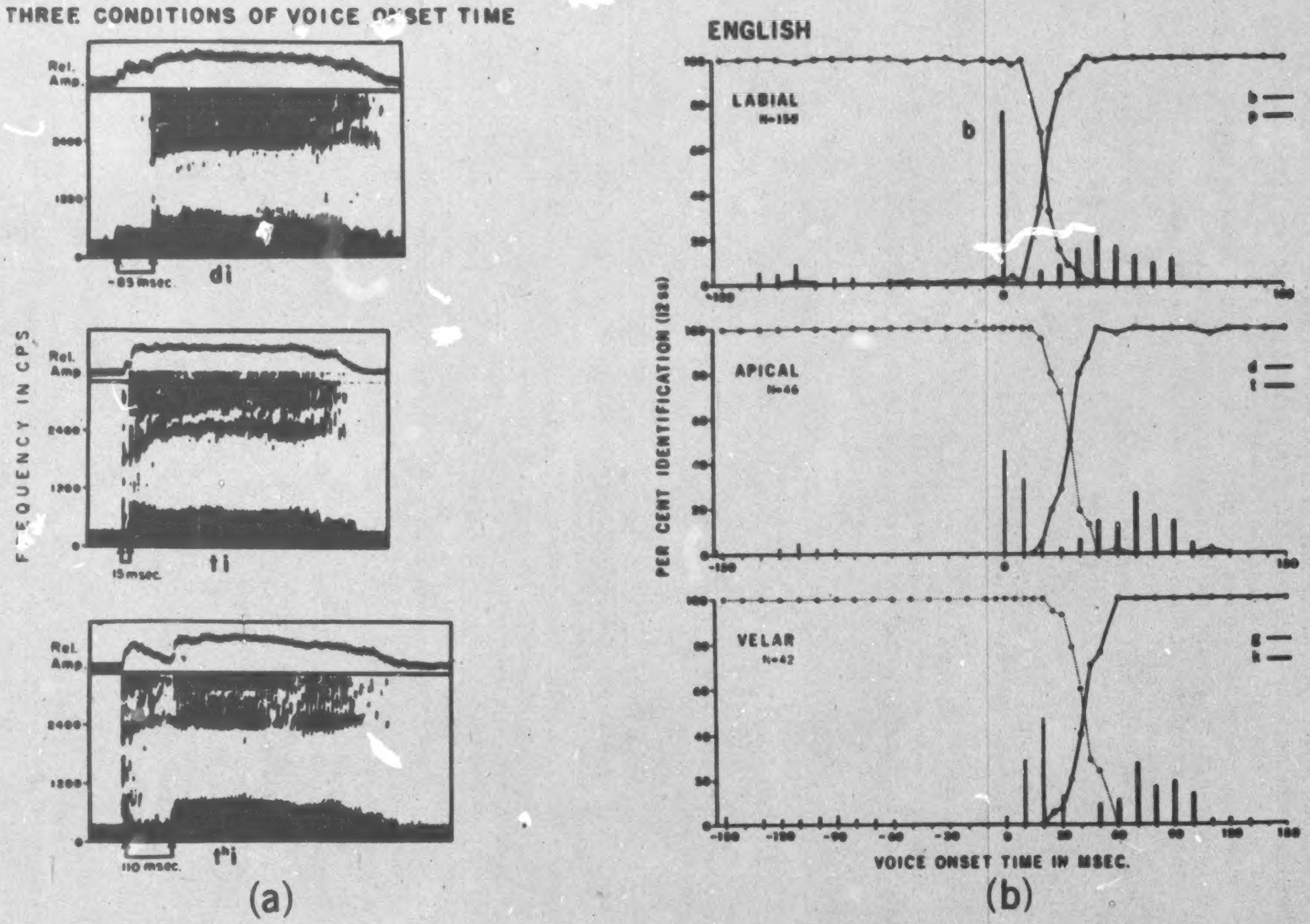

Figure 2: The left-hand panel illustrates three conditions of VOT found in Thai stops. The right-hand panel shows the distribution of VOTs for the two categories of English stops in prestressed word-initial position for the three different places of articulation. The superimposed identification functions are for a synthetic VOT continuum. The perceptual boundary corresponds well with the production boundary (Figure 1, Lisker and Abramson, 1964; and Figure 2, Lisker and Abramson, 1967a; reproduced with premission of publisher and authors).

either before or after closure can influence whether a consonant is heard as coming after the first vowel or before the second vowel in a disyllable such as /ipi/. Malmberg's results have been amplified by Darwin and Brady (1975). They examine the cues underlying the distinction between "I made rye" and "I may dry," and find that provided the stop closure is quite short, appreciable formant transitions can occur after closure without shifting the juncture to before the /d/. The perceptual system here takes into account coarticulation effects across word boundaries that are also stop closures. Although careful listeners can reliably distinguish consonant clusters with different syllable boundaries, it is not clear how important this allophonic variation is in running speech. We do not know for example how much less intelligible running synthetic speech is that puts syllable boundaries in inappropriate positions.

AUDITORY GROUPING AND FEATURE EXTRACTION

How do we group together sounds that are to be analyzed as a single source? The extensive literature on auditory selective attention has identified a number of variables that contribute to our remarkable ability to listen to one voice among many. The best known are location and pitch.

Pitch

Fletcher (1929:196) was the first to show that the ears will fuse together sounds from different parts of the spectrum when they originate from a common source. Broadbent and Ladefoged (1957) amplified Fletcher's finding and showed that fusion will occur when the sounds at the two ears have the same pitch; if, on the other hand, the signals that two ears receive are amplitude modulated at different rates, then two sources will be heard, one at each ear. The importance of this phenomenon in speech is clearly that this effect provides a mechanism whereby the different formants from a particular speaker can be grouped together as a separate perceptual channel from those belonging to other speakers, who will in general be speaking at a different pitch (Broadbent and Ladefoged, 1957). The breakdown of this mechanism can be noticed both in the concert hall when two different instruments play at the same pitch and the resultant timbre sounds like neither (Broadbent and Ladefoged, 1957), and more esoterically when, in dichotic listening experiments, two different synthetic speech sounds with the same pitch are led one to each ear. Here the impression is of a single sound, but curiously subjects tend to report the sound on the right ear more accurately, indicating that the autonomy of the two sounds is to some extent preserved (Shankweiler and Studdert-Kennedy, 1967; Darwin, 1969). A demonstration of this autonomy-preserving fusion comes from Rand (1974) who played to one ear of each of his subjects a stop-vowel syllable from which the second-formant transition had been removed. This transition, which provided the only cue to place of articulation (/b/, /d/, or /g/), was led to the opposite ear but on the same pitch and in the correct temporal relation to the main stimulus. He found that subjects could distinguish the place of articulation of the consonant and hear an additional nonspeech noise in the ear receiving the transition.

Grouping by a common pitch may be a special case of a more general phenomenon of grouping together sounds by time of onset (with each laryngeal pulse marking a new event), since the normal fusion of two sounds with a common pitch can be overridden if the two components start at different times. For example, if we first listen to formant 1 of a vowel and then add formant 2, both formants can be heard in the resultant timbre. This is not true of the vowel that can be heard by starting both formants simultaneously.[2] The concert hall again furnishes a further illustration. Turning on a radio in the middle of a sustained orchestral chord yields a timbre that is not decomposable into its component parts, but that separates out as soon as instruments change notes at different pitches.

Pitch plays a further role in auditory grouping at a more complex level. Bregman and Campbell (1971) have recently shown, although the principle was well-known to baroque composers, that when a sequence of six notes, of which three are high in pitch and three low, is played rapidly, the impression is of two separate tunes, one high and one low. This impression only occurs for fast presentation rates (the crucial rate depending on the pitch intervals employed) and is objectively confirmed by subjects being unable to make judgment *between* tunes of which note preceded which, although this ability is good *within* tunes for notes the same time apart. This effect might derive from a speech perception mechanism, enabling a listener to remain listening to a particular voice over periods of silence or pitchlessness in competition with other voices at different pitches. A recent unpublished experiment by myself and Davina Simmonds supports this idea. Simmonds asked her subjects to shadow a passage of prose presented to one ear with instructions to ignore what was presented to the other ear. Unknown to the subjects, at some time during the passage the one they were supposed to be shadowing might change to the opposite ear giving a semantic discontinuity on the shadowed ear. Treisman (1960) had previously shown that in this situation subjects occasionally made errors in which they continued to shadow the same passage after it had switched to the "wrong" ear. Simmonds's contribution was to show that whether these errors occurred or not depended on whether the intonation pattern switched between the ears. If the passages were prepared by being read continuously so that intonation was continuous across the semantic break, few intrusion errors occurred from the opposite ear, although subjects tended to hesitate. The intrusions did occur, though, when the intonation switched ears. Moreover, intrusions still happened even when there was no semantic break. Continuity of intonation thus seems to be an important factor in determining which part of the auditory input is to be treated as belonging to the currently attended channel. It remains to be seen how much this is due to short-term effects, such as those found by Bregman and Campbell (1971) and how much it is due to predictions of the expected prosodic pattern based on more complex aspects of the preceding input.

This extensive use of pitch in grouping auditory elements together suggests that it might be extracted by a different mechanism from that which is concerned with extracting information about the formant structure or timbre. This seems to be the case. Many speech signals have little or no energy at the fundamental (corresponding to the rate of vocal cord vibration) and yet the pitch is clearly heard. This problem, the missing fundamental, has indicated the need for a pitch mechanism other than the detection of place of excitation on the basilar membrane. Licklider (1951) first suggested autocorrelation as a possible mechanism for extracting this "periodicity pitch" and his idea has recently been revived by Wightman (1973), who claims that autocorrelation can handle a number of strange effects previously rather hard to explain (see Small, 1970). Autocorrelation involves quite simply a comparison (correlation) of the signal with itself delayed by varying amounts. The correlation will be maximum when one signal is delayed relative to the other by an amount equal to the periodicity of the waveform. Thinking of autocorrelation in this way leads to a realization of a possible mechanism in terms of a neural delay line (Licklider, 1951), but another mechanism that is mathematically equivalent is based on the observation that periodic signals have spectra with periodic peaks, the spacing between the peaks being related to the periodicity. Thus a device for recognizing regular patterns of excitation along the basilar membrane rather after the manner of spatial Fourier analyzers in the visual system (Campbell and Robson, 1968) could also achieve autocorrelation. This latter type of mechanism is favored by Wilson (1973) and seems to have the advantage of achieving a first stage toward the required grouping of the individual components, prior to analysis of the formant structure. Autocorrelation has also been used as the basis for automatic pitch extraction devices that are less prone to errors (such as jumping up an octave) than traditional pitch meters, which operate on a principle closer to "place" theories (e.g., Lukatela, 1973).

---

[2] Alvin Liberman: personal communication.

Location

It has been known for some time that angular separation of auditory sources helps selective attention to one rather than the other (Broadbent, 1958) provided that other cues such as pitch do not override the usefulness of location. It appears also that localization is important in determining the effect that one sound can exert on another preceding it. If a consonant-vowel syllable is played to one ear followed (say 60 msec later) by another syllable, differing in the consonant, to the other ear, then the second consonant will be reported more accurately than the first (Studdert-Kennedy, Shankweiler, and Schulman, 1970). The question now arises, if one sound masks its predecessor, how do we ever perceive a continuous flow of speech? And, since the effect also occurs in nonspeech tasks (Darwin, 1971b), how do we perceive any rapidly changing sound? The answer perhaps lies in the observation that these masking effects are less easily obtained when the two sounds used come from the same location (Porter, 1971), when any masking that does occur tends to be more symmetrical, with less predominance of backward over forward. Perhaps then, the auditory system is using location as a heuristic to decide whether two sounds are to be treated as part of the same gestalt or whether they should be distinguished and the processing of the first discarded in favor of the second.

So far there has been no indication that any of the mechanisms outlined exist exclusively for speech even though one might argue that they have arisen because of the needs of speech perception. This is not altogether surprising, but this distinction becomes more important when considering subsequent stages of analysis.

Adaptation

The discovery of single cortical cells selectively sensitive to simple properties of a visual (Hubel and Wiesel, 1962) or an auditory (Evans and Whitfield, 1964) stimulus lent physiological credence to theories of pattern recognition that suggested that the organism first detects basic stimulus properties of a perceptual object and subsequently uses this information as a basis on which to construct a percept (Selfridge and Neisser, 1960). It also renewed interest in perceptual aftereffects since some of these perceptual distortions could be neatly explained by appealing to the adaptation of detector units similar to those discovered electrophysiologically. The basic methodological axiom here is that repeated exposure to a particular stimulus weakens the subsequent response of detectors that have responded to that stimulus. This weakening then causes a distortion in the subsequent perception of any stimulus that would normally be capable of exciting those detectors. The degree to which the perception of different test stimuli is affected by previous exposure to some other stimulus thus gives an indication of what types of properties are being detected by the sensory system (e.g., Blakemore and Campbell, 1969).

This approach has been applied recently to auditory perception. Kay and Matthews (1972) have found evidence in adaptation experiments for detectors sensitive to a tone that is frequency modulated at a particular rate, thus providing one auditory analog to the suggestion by Blakemore and Campbell (1969) that the visual system contains detectors sensitive to luminance that is amplitude modulated at particular spatial frequencies. Experiments on adaptation to speech sounds have multiplied rapidly since a seminal experiment by Eimas and Corbit (1973).

This study used stop consonants differing in voicing and with one of two different places of articulation. For each place of articulation they constructed a continuum of sounds varying in VOT. Their subjects first identified isolated sounds from these two continua, then they adapted to a token of /b/, /p/, /d/, or /t/ taken from the appropriate end of the VOT continuum by listening to it 150 times in two minutes. Their perception of the two VOT continua was then retested in a series of trials each of which involved listening to a further 75 presentations of the adapting stimulus (in one minute) followed immediately by a test stimulus. The results showed that irrespective of the place of articulation of the test and adapting stimuli, the voicing boundary moved toward the adapting stimulus, a slightly greater shift being found for voiceless than for voiced adaptors. In a subsequent paper, Eimas, Cooper, and Corbit (1973) showed that this shift in the voicing boundary persists if the adapting and test stimuli are led to different ears. The effect, then, is central rather than peripheral, but what type of detector is responsible? Is it a linguistic feature detector specific to speech, or is it an acoustic feature detector that can also subserve nonspeech distinctions? Eimas and his colleagues tackled this question by using as an adapting stimulus a voiced stop-vowel syllable with all but the initial 50 msec removed. This stimulus preserves the information that voicing starts at the beginning of the sound but does not sound like speech--"just a noise" in the words of the subjects. As predicted by the linguistic feature detector notion, adaptation to this sound gave no significant shift in the voicing boundary even when the subjects were instructed to hear the sound as speech. However, Cooper (1975), in a review of the adaptation work, cites an unpublished study of Ades who does find some adaptation in this case.

The question of whether the adaptation effects obtained were due to the adaptation of a complete linguistic feature or to adaptation of the particular cues that can subserve it has been pursued by Bailey (1973). He used a linguistic dimension with well-understood multiple cues. Place of articulation for voiced stop consonants can be cued by the second- and third-formant transitions. Bailey first showed that the adaptation effect could not be occurring at the level of a detector responding to place of articulation per se, since when subjects adapted to the syllable /be/ they showed more shift along the /be/-/de/ continuum than along the /ba/-/da/ continuum (Figure 3 a, b; see also Ades, 1974a). They did show some shift in the latter case, and Bailey's second experiment

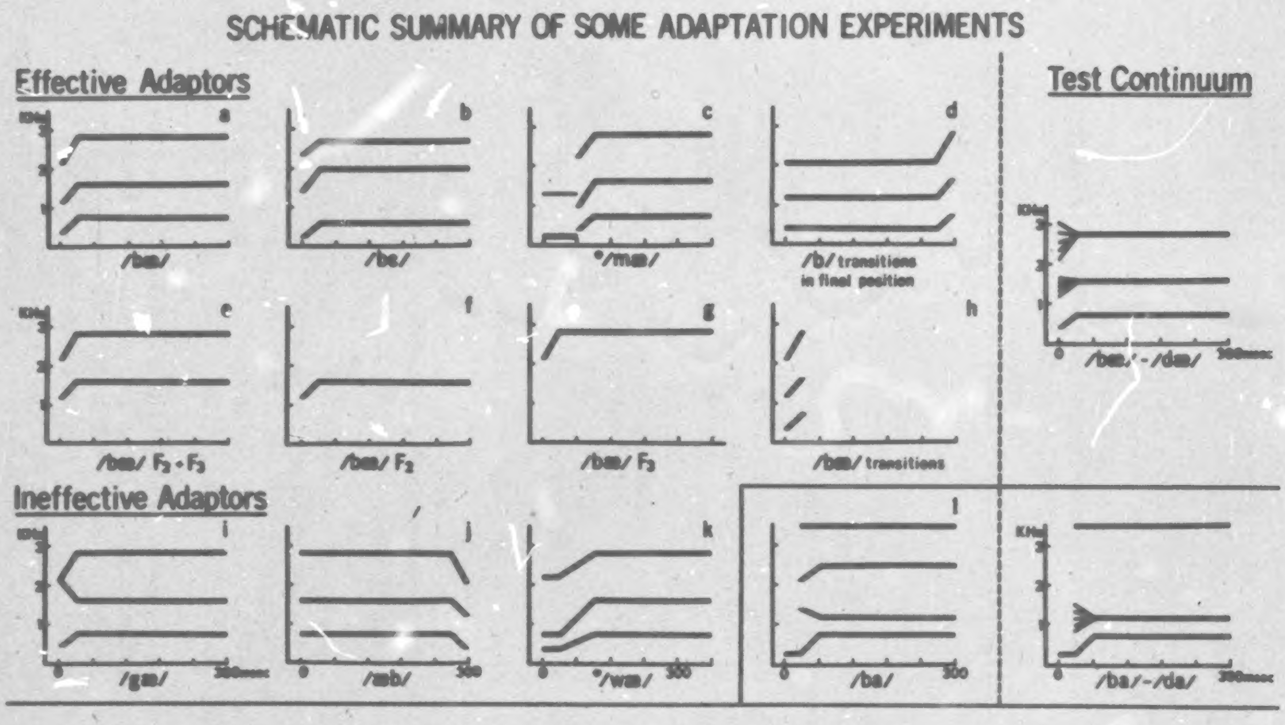
SCHEMATIC SUMMARY OF SOME ADAPTATION EXPERIMENTS
Effective Adaptors
Test Continuum
Ineffective Adaptors
KHz
1
2
3
a
/bæ/
b
/be/
c
*/mæ/
d
/b/ transitions in final position
e
/bæ/ F2+F3
f
/bæ/ F2
g
/bæ/ F3
h
/bæ/ transitions
/bæ/-/dæ/
i
/gæ/
j
/æb/
k
*/wæ/
l
/ba/
/ba/-/da/
0
300msec
300

Figure 3: Adapting with any of the "effective adaptors" gives a significant shift in the boundary between /bæ/ and /dæ/ along the test continuum shown on the upper right-hand side, so that more sounds are heard as /dæ/ after adaptation than before. This shift is considerably greater for the original /bæ/ than for the others. Adapting with any of the first three "ineffective adaptors" (i,j,k) gives no significant shift in the /bæ/-/dæ/ boundary along the same test continuum. In addition, adapting with the last of the "ineffective adaptors" (which has an F_2 transition on the /b/-/d/ boundary) gives no shift in the lower test continuum (which has no F_3). The two spectrograms marked with an asterisk (c,k) illustrate experiments that used natural speech adaptors. The other spectrograms are schematic illustrations of synthetic speech adaptors. The experiments illustrated in this figure are all referred to in the text.

FIGURE 3

suggested why. Here, rather than changing the subsequent vowel, the particular formant transition used to cue the place distinction was changed. One set of stimuli used formants 1, 2, and 4 (F_4 had no transition) with the place distinction being cued only by changes in the second formant. The other set of stimuli used all four formants but with a neutral F_2 transition so that the place distinction was being cued by changes in the third formant. Adaptation had a significant effect on the phoneme boundary when both adapter and test items were from the same set. However, when they came from different sets there was only an effect when the adapting stimuli were distinguished by a change in a formant that was present, if constant, in the test stimulus. When the adapting stimulus varied in F_3 but the test stimulus had no F_3, then no adaptation effect was observed (Figure 3 l). Here there could be no processing of F_3 transitions to reveal the underlying distortion of such processing by the adapting sequence. In a subsequent experiment Bailey showed that some effect of adaptation returned if the test stimulus was given a flat third formant.

These results give clear evidence that adaptation can occur at the level of detectors for specific acoustic features that precede the subsequent pooling of these features for a decision about the linguistic dimension. This conclusion has been confirmed by a number of recent studies. The boundary for place of articulation for stop consonants can be shifted by presentation of the isolated second and third formants (Figure 3 e, f, g; see also Tartter, 1975), by the isolated formant transitions (Figure 3 h; Ades, 1973; Tartter, 1975), or by the second- and third-formant transitions accompanied by an inappropriate, nonspeechlike first formant transition (Tartter, 1975). Provided, then, that the adapting stimulus contains cues that distinguish the items on the test continuum, the adapting stimuli themselves need not be heard as speech sounds, as pointed out by Cooper (1975).

Adaptation of specific acoustic cues can also explain why adaptation fails to generalize between initial and final consonants. Adapting to /bæ/ will shift the /bæ-dæ/ boundary but will have no effect on the /æb-æd/ boundary, and vice versa (Figure 3 j; Ades, 1974a). The reason for this is that for the synthetic speech sounds used in this experiment the formant transitions that cued the initial stop consonant were mirror images of those that cued the final stops and so were presumably served by different detectors. This explanation has been strengthened by an experiment by Tash (1974). Tash placed the formant transitions appropriate for initial stops at the end of a steady vowellike sound whose formant values were the same as the *start* of the transitions. This gave nonspeech sounds with the transitions in final position (Figure 3 d). These sounds *were* effective at shifting the boundary for initial stops. Thus, provided the direction of the transitions is preserved, adaptation can occur between initial and final positions.

Adaptation at the auditory feature level, though undoubtedly present, is not the whole story. Stop consonants, differing only in place of articulation, can be synthesized to have identical first-formant transitions so that they are distinguished only by the second- and third-formant transitions. If adaptation were occurring only at the auditory feature level, we might expect that the presence of the (constant) first-formant transition would be irrelevant to the adaptation effect of place of articulation, since it does not carry any distinguishing information. However, it is clear that although the isolated second and third formants or their isolated transitions do shift the place-of-articulation boundary, this effect is much less than if the first formant is included (Ades, 1973; Tartter, 1975).

To explain this dependency we need to consider levels above the auditory feature. Let us assume that a number of auditory feature detectors map in a hierarchical way onto some higher-level unit. For example, three rising formant detectors might map onto a detector for labial place of articulation or onto a particular syllable; at least one of these, the rising first-formant detector, will presumably also map onto units at this higher level that have different places of articulation.

After adapting to a complete speech sound with all three formants present, all three of the auditory feature detectors will have been fatigued, but after adapting to the sound without the first formant, only the second- and third-formant detectors will have adapted. There are now two ways to explain the greater boundary shift with the complete stimulus. We can assume that through some nonlinearity in the system the adapted first formant leads to a greater reduction in the firing of the higher-level unit already receiving a reduced input from other (adapted) detectors than in the firing of a unit receiving unadapted input. Alternatively, following Tartter (1975), we can assume that some adaptation is occurring in the higher-level unit itself. Although on the basis of the data presented so far, the former hypothesis is perhaps more economical, additional adaptation at a higher level is made more likely by two types of study: split formants and cross modality. In the split formants experiment the first formant of a syllable is led to one ear while the other two are led to the opposite ear. If the two sounds are played simultaneously, then the entire syllable is heard, but if they are played one after the other, then two nonspeech sounds are heard. In an ingenious experiment, Ades (1974b) showed that there was greater adaptation when the sounds were presented simultaneously than when they were played at different times. Playing the components at different times should have no differential effect on the adaptation to the detectors of those components, but it would prevent adaptation to the higher-level category since that is never heard. The greater adaptation for simultaneous presentation thus argues for adaptation also occurring at some level above the auditory feature level. The second type of evidence comes from cross-modal adaptation. Repeated visual presentation of a syllable the subject reads silently gives a shift in the boundary for voicing (Cooper, 1975), which is specific to position within the syllable (Eimas, in preparation) so that a visually presented "bæ" will adapt the auditory /bæ-dæ/ boundary but not the /æb-æd/ boundary.

What then might be the level of this higher category? We can put forward three candidates: the phonetic feature, the phoneme, and the syllable. A number of authors have proposed that adaptation effects can occur at the phonetic feature level (Ades, 1974a; Cooper, 1974; Tartter, 1975), the argument being that there is some generalization for adaptation to place of articulation across vowels (Ades, 1974a; Cooper, 1974; Tartter, 1974) and across manner of articulation (Cooper and Blumstein, 1974). But these arguments are not strong since this generalization can perhaps be handled by adaptation at the auditory level (see Cooper, 1975, for a discussion of these issues). Cooper and Blumstein find significant shifts in the /bæ-dæ/ boundary after adaptation to /bæ, mæ, væ/ (Figure 3 c) but very little after adaptation to /wæ/ (Figure 3 k). Since all these sounds except /wæ/ contain similar formant transitions, whereas /wæ/ has slower transitions, the observed adaptation effects can be explained in terms of rate-specific formant transition detectors. Rather more difficult to explain are the cross-vowel adaptation effects, but these can perhaps also be handled at the auditory level. Both Tartter (1975) and Ades (1974a) have shown that adapting to a stop consonant in front of one vowel gives some shift in the boundary for place of articulation of a consonant in front of another vowel. Both of them used three formant syllables with the third formant common to both vowels. Since third-formant transitions are cues to place of articulation, it is not surprising that some cross adaptation occurred.

Indeed, in Tartter's experiments it is possible to compare directly the extra adaptation effect attributable to the syllables sharing a common phoneme while controlling for common acoustic cues. Tartter finds that the third formant from a /bæ/ produces a significant shift in the /bæ-dæ/ and the /dæ-gæ/ boundaries. This shift is in opposite directions for the two boundaries, since /bæ/ and /gæ/ both have rising third formants, while /dæ/ has a falling third formant. However, Tartter also examines the shift in these boundaries after adaptation to /bi/ and /gi/. These stimuli are of interest since they both have the same third formant as /bæ/, but different first and second formants. On purely auditory grounds then we would expect /bi/ to have the same effect on a /bæ-dæ/ boundary as the /bæ/ third formant, while if some adaptation were occurring at the feature or phoneme level, /bi/ should have a larger effect on the /bæ-dæ/ boundary than the isolated third formant. A similar prediction can be made *mutatis mutandis* for /gi/. In fact, the average adaptation effect when the phoneme is shared is only 20 percent greater than when auditory factors alone are contributing to the adaptation.

If adaptation is occurring at some level beyond the auditory feature, then it would appear that the most appropriate level, at least for stops cued by formant transitions, is the *syllable*. This would explain why an entire speech sound is a more effective adaptor than the acoustic discriminanda alone and also why *cross-modal* adaptation is specific to initial or final position. Assuming that some adaptation occurs at both an auditory and a higher level also allows one to explain why the /bæ-dæ/ boundary does not shift after adaptation to /gæ/ despite significant shifts being found after adaptation to the third formant that /bæ/ and /gæ/ have in common (Cooper, 1974; Tartter, 1975).

There are a number of lines of evidence to suggest that auditory feature detectors are specific to a sound's spatial location. Ades (1974b) found that he could adapt the two ears simultaneously to different sounds using dichotic presentation. Thus, one ear received /bæ/ at the same time as the other ear heard /dæ/. The direction of the shift in the /bæ-dæ/ boundary was different for the two ears. Ades also showed that after adapting to a sound presented to one ear alone, there was incomplete (55 percent) transfer to the opposite ear. Although Ades interpreted these results in terms of peripheral versus central adaptation, it would be equally valid to interpret them in terms of location-specific auditory detectors. Recent support for this notion comes from an experiment on the verbal transformation effect (Warren, 1968). Repeated listening to a word causes it to lose its meaning and change its sound so that it is perceptually transformed into another word. It seems likely that at least part of this effect is due to adaptation at the acoustic level (see Lackner and Goldstein, 1974). Warren and Ackroft (1974) have shown recently that if the same word is presented to the two ears but slightly offset in time so that two distinct utterances are heard, then different transformations can be heard in the two ears. This suggests again that there are different sets of detectors for different spatial locations capable of being differentially adapted.

Such wanton proliferation of auditory detectors might seem unwarranted and unnecessary, but without multiple detectors for identical sounds it is difficult to see how two separate streams of speech could be handled at the same time. That this does appear to be the case is shown by studies of selective attention. If a subject has to shadow a prose passage read to one ear while at the same time trying to make a manual response whenever a target word is played into either ear, he will fail to detect the vast majority of targets on the unattended ear, while successfully responding to those on the shadowed ear (Treisman and Riley, 1969). However, if the subject is conditioned before the experiment by being subjected to an electric shock each time he hears a word belonging to a particular semantic category and subsequently has to respond manually to words belonging to that semantic category while shadowing, then although virtually none of the words on the unattended ear produce a manual response, over a third give a galvanic skin response (Corteen and Wood, 1972). Thus, although the semantic properties of words on an unattended channel rarely reach consciousness, it can be shown that their semantic properties have been extracted. Some basic perceptual processing can occur for different speech streams at the same time.

In summary, then, the work on adaptation gives good evidence for detectors tuned to complex auditory patterns, such as particular formant transitions that may exist in multiple sets, each set being maximally responsive to sounds from a particular location. The adaptation work gives evidence also for units at a more complex level than the auditory feature. While it is not yet clear what level these additional units are at, it is suggested that the available evidence is not incompatible with formant transition information being mapped directly onto a syllabic unit. For more discussion of these points and also a review of the adaptation work on voicing, see Cooper (1975).

As a cautionary footnote to the work on adaptation, it is possible that some of the phoneme boundary shifts found in adaptation experiments may be due to factors other than the adaptation of various detectors. In particular, it is possible that some of the observed effects can better be looked on as criterion shifts brought about by a change in the range of stimuli recently presented to the subject. Brady and Darwin (in preparation) have found that when subjects have to classify stop consonants presented in blocks of 16 trials during which all the stimuli come from a subrange of the voicing continuum, the phoneme boundary moves as a function of the location of the subrange used in that block (Figure 4) and the subrange used in the preceding block. This is true whether the subjects have heard sounds from the entire range at the beginning of the experiment or not (although the effects are larger when they have have not). The direction of the shift is such that sounds near the boundary are heard as being more voiceless when they occur in a range that extends toward the voiced end. However, it is unlikely that these range effects can themselves be explained by adaptation since Sawusch, Pisoni, and Cutting (1974) have failed to find a similar shift in the voicing boundary when they varied the probability distribution of the stimuli presented rather than their range. These authors used two different probability distributions, one in which the most voiced stimulus was four times as likely to occur as any of the other stimuli to be identified and another in which the most voiceless stimulus was the more probable. Their experimental design differs from our range experiment in that it always includes some tokens from each end of the range. Since the number of these tokens is very small, it seems unlikely that they could be causing much of a change in the adaptation state of property detectors, rather we should perhaps conclude that they are sufficient to cause a shift in some criterion setting used to evaluate the phonetic significance of the output of the property detectors.

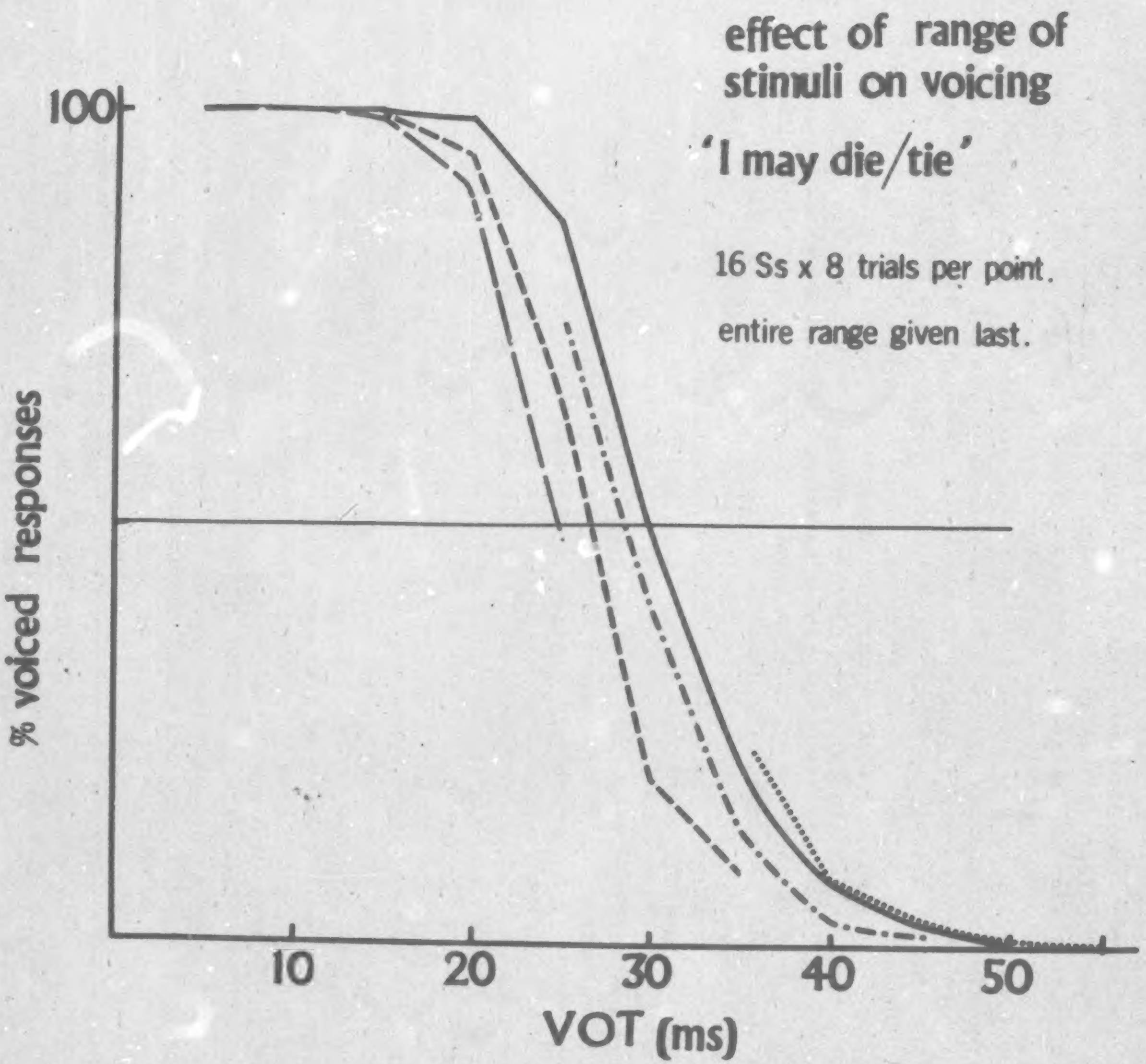

Figure 4: Identification functions for stops differing in voice onset time, presented to subjects, in blocks of trials confined to a particular range of voice onset times. Each ogival segment extends over the range of sounds used in a particular block and the solid line shows subjects' responses to stimuli in a block that covered the whole range. There is a significant change in the voicing boundary with changing range.

FIGURE 4

Dichotic Masking

Although less well-developed than the work on adaptation, some experiments on dichotic masking are compatible with the idea that the extraction of acoustic features can be interrupted by the subsequent presentation of another sound that shares similar features.

The initial impetus to this work came from the previously mentioned experiment by Studdert-Kennedy, Shankweiler, and Schulman (1970), which showed that when two different stop-consonant-vowel syllables were presented one to each ear with a temporal offset between them, the second syllable tended to be reported more accurately than the first, over a range of offsets from 15 to 120 msec. Although initial interpretation of this effect was in terms of the interruption of some special speech processing device, on the grounds that vowel sounds were less prone to this masking (Porter, Shankweiler, and Liberman, 1969), it now seems likely that the effect arises at the acoustic feature extraction stage after the initial grouping process. The reasons for this are first that the vowel/consonant dichotomy seems irrelevant, since Allen and Haggard (1973) have shown, confirming a prediction by Darwin and Baddeley (1974), that acoustically similar vowels suffer greater backward than forward masking, whereas acoustically different vowels do not. Second, greater backward than forward masking occurs for sounds distinguished by rapid transitions, irrespective of whether these transitions cue a linguistic distinction or not (Darwin, 1971b). Third, Porter and others[3] have found that the stop consonant in a CV syllable can be masked by subsequent contralateral presentation of the syllable's second formant; again, this shows greater backward than forward masking.

Reasons for supposing that these masking effects are due to the interruption of the extraction of some acoustic feature rest on analogies with visual masking. Here it is possible to distinguish between two types of masking: integration and interruption. Integration masking, which is at a prior stage to interruption masking, is more evident if target and mask are presented to the same eye; it then depends mainly on the relative energies of target and mask and to a much smaller extent on the contour relationship between the two. On the other hand, interruption masking occurs equally whether target and mask are presented to the same or to different eyes, and appears to be independent of the relative energies of the target and mask, but does require that the two share similar contours (Turvey, 1973). Darwin (1971b) pursued the analogy between auditory and visual masking by distinguishing between the sets of sounds used for the target and the mask. Previous experiments had confounded the two by drawing both target and mask from the same set of sounds, and then asking the subject to report either both sounds or the one on a particular ear (Kirstein, 1970). Darwin used the syllables /be, pe, de, te/ as a target set and one of four masks (/ge, e, ɔ/ and nonspeech steady-state timbre) presented on the opposite ear either 60 msec in front of or behind the target. He found that for place of articulation the amount of forward masking was the same for all the masks and rather small, on the other hand the amount of backward masking did depend on which mask was used, being very much greater for the /ge/ mask than for the others, which showed only minimally greater backward than forward masking. The perception of voicing showed only slightly greater backward than forward masking and this was the same for all the masks used. As in vision, then, there is greater backward over forward masking only when the mask shares some features with the target.

---

[3] Personal communication.

The question of what type of features the target and mask must share has been taken up by Pisoni and McNabb (1974; Pisoni, 1975). Their experiment uses as a target set /ba, da, pa, ta/, and six different masks /ga, ka, gæ, kæ, ge, ke/. They too find that the amount of backward masking depends on the similarity between the target and the mask, backward masking is much greater when the mask has the same vowel as the targets.

This result is compatible with the idea that backward masking is sensitive to the particular auditory features present in the consonant rather than to its phonetic features, since only the former vary when the vowel is changed. However, in a subsequent experiment they show that similar results also apply in forward masking, when the mask precedes the target. This result causes them to interpret both their experiments in terms of an integration hypothesis, since it is crucial for the interruption hypothesis that backward masking should be greater than forward masking, but it is possible to accommodate the effect of target-mask similarity into an integration theory. This interpretation is appropriate for Pisoni and McNabb's (1974) results (although their data do show a slightly greater effect of backward masking) but it cannot be used as Pisoni (1975) claims as an interpretation of those results that have shown appreciably greater backward than forward masking.

How can we reconcile these results? It is clear from the experiments that do show greater backward than forward masking that some forward masking is occurring (Studdert-Kennedy et al., 1970), so it may not be inappropriate to suggest that some integration masking is present. Indeed this would not be surprising since the sounds used in all these experiments have a common pitch and so are likely to fuse together to some extent, depending on their onset asynchronies. The important question is why some experiments show much more backward than forward masking (Kirstein, 1970; Darwin, 1971b; Porter, 1971; Studdert-Kennedy et al., 1970) whereas others do not (Pisoni and McNabb, 1974). The answer probably lies in the different tasks required of the subject. Pisoni and McNabb's experiment is unique in that the subject always knew onto which ear the target would come and whether it would precede or follow the mask; whereas in all the experiments that have shown greater backward than forward masking, the subject was in doubt either as to the ear on which the target sound would arrive or as to whether it would be the first or the second. The presence of both these physical cues may then allow the subject to stop the second sound from interrupting the processing of the first.

In summary, the data on dichotic masking of speech sounds suggest that both integration and interruption processes are occurring and that in both types the similarity of the target and the mask can affect the amount of masking obtained. However, in circumstances where interruption masking is clearly occurring (backward masking greater than forward) the evidence suggests that it is occurring at the level of auditory features. Little more can be inferred at present from these experiments about the nature of these features. It is also unlikely that the masking experiments will provide as direct an access to them as adaptation experiments have, both because of the different types of masking that may occur and because of the contribution that other factors such as echoic memory may play in performance on masking experiments (Allen and Haggard, 1973; Darwin and Baddeley, 1974). However, the ease with which backward masking effects are shown for rapid acoustic transitions, compared with the unreliable and small effects for vowels and pure tones (Massaro, 1970; Pisoni, 1972; Cudahy and Leshowitz, 1974), suggests that the auditory features are likely to be complex or time-varying.

Objective methods then are beginning to provide a way of defining psychologically important auditory features. An interesting question is whether these will turn out to be the same as those that seem to be important for other reasons. Fant (1964) stresses the importance of looking in the speech signal for "auditory patterns" that might serve either as a direct indication of a segment's identity (as in /s/) or, more usually, would provide the raw materials for the more complex processes required to interpret them in terms of linguistic categories.

MODELS OF SPEECH PERCEPTION

We are now in a position to see the magnitude of the problem posed by speech perception, even at the phonemic level. Because of effects such as coarticulation the speech signal resists any effort to segment it into acoustically defined portions that are influenced only by a particular phone, except in a very restricted set of cases. Some segmentation is possible according to purely acoustic criteria (Fant, 1967) and we have seen that there is growing evidence that auditory features are extracted as part of the perceptual process. But where do we go from here? What type of process mediates between the auditory feature and the phonetic category?

Formant transitions do not provide a simple invariant cue of the form "a slightly rising transition lasting 40 ms" (Cooper et al., 1952), but can we say that they do not provide a set of invariant exclusive disjunctions of the form "a falling second formant lasting 40 msec ending around 800 Hz OR a slightly rising second formant lasting 40 msec ending around 2700 Hz, etc.?" Rand (1971) has provided a simple demonstration that this type of invariance is not sufficient. He constructed two sets of synthetic syllables whose vowels had the same second formant but different first formants, in one case the vowel sounded like an /æ/ produced by a child, and in the other, an /ɛ/ produced by an adult. Each of these vowels was preceded by formant transitions to give the three stops /b, d,g/. Rand found that the best transition for a /d/ response varied as a function of the apparent vocal-tract size. The significance of a particular second-formant transition thus depends on the interpretation of the formant pattern as a whole, rather than on the value of the formant that it leads to.

The main dichotomy in models of speech perception has been between active (analysis-by-synthesis) and passive models. The distinction between these two types of model is that while the passive model sees the primary categorization process as being due to some filter network, whose decision criteria change relatively slowly with time (Morton and Broadbent, 1967), the active model sees it as being the result of matching the input signal to an internally generated representation that can change rapidly relative to that signal (Liberman et al., 1962; Stevens and House, 1972).

Active Models

One of the motivations behind active models of speech perception (e.g., Liberman et al., 1962; Stevens and House, 1972) is that of economy. Given that the processes of speech production and perception are both highly complex and formally similar, would it not be an economical solution to combine the two? Economy of description, though a fundamental criterion for the linguist, is not an infallible guide for the psychologist, and indeed there is no shortage of evidence for the independent operation of perceptual and production mechanisms. Listening to speech while speaking oneself is commonplace but is not readily explained away by active models. The more extreme example of simultaneous translation from one language to another (Neisser, 1967:217) illustrates this point well; here the very language of perception and production are different. Other examples of independence appear from the clinic. Patients with the two hemispheres separated by cutting the corpus callosum can show by activities of the left hand that they comprehend instructions presented to the right hemisphere, but they cannot show this by speaking (Gazzaniga and Sperry, 1967). Congenitally anarthric individuals appear to have normal speech perceptual abilities (Lenneberg, 1962), and children, who by virtue of an articulatory defect are unable to make a particular articulatory distinction, show no corresponding perceptual impairment (Haggard, Corrigall, and Legg, 1971). Experiments such as these suggest that the ability to perceive speech comes through "the distinctiveness of the speech wave which we have acquired by being exposed to language in the first place and by reference to our own speech only in the second place" (Fant, 1967:113).

Combining the mechanisms of production and perception also offers no way of accounting for variables that change between speakers. That interspeaker variation is a significant perceptual problem is indicated both by experiments on speaker normalization (Ladefoged and Broadbent, 1957) and by the impaired performance of automatic speech recognition devices when tested on more than one speaker (see Hyde, 1972, for a review).

Although the problem of the lack of acoustic/phonetic invariance is cited as one reason why an active model is needed, it is not clear that such a model will solve the problem. The trouble lies in deciding what is to be compared with what. The acoustic signal is presumably represented in terms of auditory parameters, while the internally generated articulatory representation is in terms of articulatory parameters. Before any direct comparison can be made there must be some translation between the two. To get round this problem, Stevens and House (1972) propose that a "catalog of relations between acoustic and articulatory instructions [of approximately syllable length] is built up in a child at an early age... As the child produces new and different articulatory movements, new items are added to the catalog" (p. 53). But if the mature speaker has this catalog, why bother with analysis-by-synthesis at all? Could you not simply look up the acoustic pattern in the catalog? Analysis-by-synthesis itself gets us no closer to solving the invariance problem.

Analysis-by-synthesis is also seen as a way of using the constraints of the language to aid perception. This is dealt with later, but, in anticipation, experimental evidence indicates that context is not used in the way suggested by active models, at least at the word level.

Both the Haskins model (Liberman et al., 1967) and the Stevens and House (1972) model emphasize that different perceptual mechanisms are needed for phonemes that do or do not have invariant acoustic cues. Stevens and House perform a preliminary analysis on the signal in order to guide the subsequent synthesis, while the Haskins model allows only the more variable or "endoded" sounds to engage the speech processing mechanism. The perceptual significance of this dimension of "encodedness" has been claimed from a range of different experimental paradigms. These experiments have shown that stop consonants (the least invariant phonemes) produce different results from vowels (the most invariant phonemes), while other phonetic categories fall somewhere in between.

Perception of Different Phonetic Categories

The earliest demonstration of this difference between phonetic classes was a set of experiments comparing categorization and discrimination of synthetic speech sounds. Since the second-formant transition can act as a sufficient cue for initial stop consonants, a continuum of sounds can be constructed varying in the extent and direction of this transition. Subjects will then label with some consistency sounds taken from this continuum as /b/, /d/, or /g/, depending on their position along it. If pairs of sounds adjacent on this continuum are then played to subjects, their ability to discriminate between the members of a pair will be rather poor unless the pair happens to straddle the boundary between sounds labeled as different phonemes. For a continuum consisting of vowels, on the other hand, discrimination is good throughout the continuum (Liberman, Harris, Hoffman, and Griffith, 1957; Fry, Abramson, Eimas, and Liberman, 1962). Other paradigms that have shown differences between stops and vowels include laterality and dichotic masking experiments. After simultaneous dichotic presentation, stop consonants are recalled more accurately from the right than from the left ear, whereas vowels show the effect less consistently (Shankweiler and Studdert-Kennedy, 1967). Similarly, if two different sounds are led one to each ear but with a temporal offset of around 60 msec, the second sound is recalled more accurately than the first for stops, but not for vowels (Studdert-Kennedy, Shankweiler, and Schulman, 1970). Other classes of speech sounds have given results in the laterality paradigm intermediate between stops and vowels. For example, place of articulation for fricatives is cued mainly by the spectrum of the friction, but intelligibility is increased if appropriate formant transitions are added (Harris, 1958). The friction is a comparatively invariant cue to place of articulation, whereas formant transitions are more variable. In keeping with the predictions for active theories, Darwin (1971a) found that place of articulation for fricatives was only reported better from the right ear if formant transitions were present. Similarly, Cutting (1972) has shown that liquids (/r,l/), which can be regarded as having an intermediate amount of invariance, show an ear difference between that for stops and vowels.

That these experimental differences should be attributed to the relative amounts of invariance or encodedness of different phonetic classes has been questioned. Fujisaki and Kawashima (1968) offered an alternative explanation of the discrimination experiments. They observed that the discrimination of short-duration vowels showed clearer peaks at the phoneme boundary than did long-duration vowels. On the basis of this evidence, they proposed that performance in a discrimination task is determined both by the categorization process and by uncategorized information held in a buffer store. Pairs of sounds that differ in terms of the categorization process can be judged different on that basis, but if they are categorized as the same phoneme, then comparison is made between their representations in the buffer store. Fujisaki and Kawashima showed that short vowels were perceived more categorically than long vowels. They suggested that a less accurate comparison could be made between the buffer store representations of stop consonants and brief vowels than between those of long vowels on account of their duration. This result has since been confirmed by Pisoni

(1971), who has also shown (Pisoni, 1973) that the accuracy of discrimination between pairs of long- and short-duration vowels decreased as the interstimulus interval increased, whereas the discrimination of pairs of stop consonants remained stable with time. There was a marked difference between the within-category discrimination scores for stop consonants and for vowels of the same duration. Clearly, cue duration of itself is not an adequate explanation of the discrepancy. One explanation of these effects (Liberman, Mattingly, and Turvey, 1972; Pisoni, 1973) is that a special mechanism responsible for the perception of stop consonants precludes the subsequent use of auditory information for nonphonetic judgments. If this explanation is valid, then Fujisaki and Kawashima's (1968) model needs to be altered to prevent auditory information being used after it has been categorized in a particular way. This explanation also renders the hypothesis of a special processing mechanism for stop consonants immune from attack along the lines proposed by Fujisaki and Kawashima.

This question of the relationship between the categorization process and the buffer memory trace from which it is derived has been examined in a different context by Darwin and Baddeley (1974). As a result of experiments on acoustic memory based on recency effects in immediate serial recall of lists of items (Crowder and Morton, 1969; Crowder, 1971), they proposed that acoustic memory is not influenced by the categorization process. Rather, it is simply an analog representation of the acoustic stimulus, which becomes degraded with the passing of time (Darwin, Turvey, and Crowder, 1972). The result of this degradation is that acoustically fine distinctions are lost before acoustically coarser ones. Darwin and Baddeley (1974) suggested that the various experiments that had purported to show differences in categorization mechanisms for different phonetic classes merely reflected differential contributions from acoustic memory because of the different acoustic confusability of items within different phonetic classes. Little useful acoustic information about place of articulation for a stop consonant could be obtained from acoustic memory a short time after its arrival, not because it has been categorized as a stop, but because it is acoustically very similar to other stops with different places of articulation. However, this information will be less useful in distinguishing between different stop consonants than will similarly degraded information, which need only be put into acoustically coarser categories.

According to this account, the reason some speech sounds show laterality effects where others do not is because the vocabulary of sounds used is sometimes sufficiently acoustically different for useful information to persist for some time in acoustic memory. Thus the left hemisphere is given more time to categorize a left-ear signal, which, by virtue of poorer neural connections to that hemisphere, is degraded compared with the right-ear signal (Darwin, 1973). Similarly, the reason some sounds show more backward masking than others is because for acoustically distinct vocabularies the categorization mechanism can use the information in acoustic memory to take a second pass at a previously interrupted categorization. This hypothesis predicts that there should be a three-way correlation between laterality, dichotic masking, and acoustic memory experiments; so that the greatest evidence of acoustic memory (in, for example, recency experiments) is given by vocabularies of sound showing the least laterality effects and the least dichotic backward masking effects. This correlation has been shown in a number of experiments.

Acoustically similar vowels (such as /ɪ,ɛ,æ/) show little evidence of useful acoustic memory in recency experiments (Darwin and Baddeley, 1974) whereas acoustically distinct vowels (such as /ɪ,æ,u/) do. Similarly, acoustically similar vowels show a significant right-ear advantage while acoustically distinct vowels do not (Godfrey, 1974), and acoustically similar vowels show more dichotic backward masking than do acoustically distinct ones (Allen and Haggard, 1973). Syllable-final consonants show more recency than syllable-initial consonants if they are acoustically distinct (/g,ʃ,m/; Darwin and Baddeley, 1974) but not if they are acoustically similar (/b,d,g/; Crowder, 1973), and likewise syllable-final consonants show more right-ear advantage than syllable-initial if they are cued by slow transitions (/r,l/; Cutting, 1972) but not if they are cued by fast ones (/b,d,g/; Darwin, 1969). In stop consonants, voicing shows less backward masking than does place of articulation (Darwin, 1971b) but more recency.[4] Adding appropriate formant transitions to fricatives makes them show a right-ear advantage for place of articulation (Darwin, 1971a) but gives no increase in the size of their recency effect (Crowder, 1973).

The success of this hypothesis rests not only on showing that the utility of auditory information depends on the acoustic similarity of the items used rather than on their phonetic class, but also on showing that under suitable circumstances auditory information is available from sounds belonging to acoustically similar categories, such as the stops. An experiment by Pisoni and Tash (1974) gives direct evidence that some auditory precategorical information is available from stop consonants. They used a same-different reaction-time paradigm (Posner and Mitchell, 1967) measuring subjects' reaction times to pairs of sounds drawn from a continuum between /b/ and /p/. Subjects had to decide whether the two sounds were the same phoneme or not. Their reaction times showed that it took them longer to decide that the two sounds were the same phoneme when the sounds were physically slightly different (but still within the same phoneme category) than when they were identical. They also found that the time to decide that the two sounds were different was faster when the sounds differed by a larger distance on the continuum than a small, even though the sounds always fell within different phoneme categories. Some precategorical information must have been available to the subjects over the half-second or so that separated the onsets of the two sounds.

It would be premature then to claim that we have any psychological evidence for different phonetic classes of sounds being perceived by different types of categorizing mechanisms. This does not mean of course that there are no differences, it merely shows that the paradigms used so far are not sensitive to what differences there may be. Deprived of this empirical support, active models become less plausible. But in rejecting the active model as a mechanism for categorization, we must be careful not to reject it as a statement of the problem. The active model is correct in emphasizing that knowledge of the vocal tract must be used in order to categorize speech, but it appears to be incorrect in suggesting that the mechanism by which this is done is an active analysis-by-synthesis. Failure to use knowledge of the mechanisms and acoustics of the vocal tract is an important reason why contemporary attempts at machine recognition of speech have been unsuccessful (see Hyde, 1972, for a review). The success of programs of analyzing visual scenes is related to the sophistication of the geometrical constraints they have employed (Guzman, 1968; Clowes, 1971; Mackworth, 1973). Perhaps we can expect a similar improvement in speech recognition with the use of more sophisticated knowledge about the vocal tract. Although this knowledge exists in a variety of forms, including programs to synthesize speech by rule (Holmes et al., 1964; Mattingly, 1968; Kuhn, 1973), it has not yet been applied systematically to perceptual problems.

---

[4] A. Thomasson: personal communication.

Computational procedures exist that allow the cross-sectional area function of the vocal tract to be estimated directly (without recursive procedures) from the acoustic waveform (Atal, 1974). The advantages to the perceptual system of performing such a transformation are clear in that many of the problems raised by coarticulation evaporate; but is it likely that this is a first stage in perception? For reasons outlined by Haggard (1971), such a process would be more likely to appear in the perception of vowels where the spectrum is simpler, than for consonants with their additional sources of acoustic energy. Haggard suggests that the acoustically complex consonants may be perceived by heuristics that map directly from acoustic features to phonetic categories (as we have seen suggested for voicing), while vowels may use a procedure that computes something like the vocal-tract area function. Evidence from vowel perception, however, indicates that if the perceptual process passes through an articulatory representation of the vowel, this is perceived heuristically, or by "rules of thumb" that do not achieve a representation as detailed as a vocal-tract area function.

Carlson, Granstrom, and Fant (1970) report the results of experiments in which subjects are asked to adjust the second formant of two-formant vowels to match vowels composed of four formants. The value of this F_2' lies between the values of the matched vowel's second and third formants (except for [i:], where it is above F_3) in a position that Carlson, Fant, and Granstrom (1973) show is predictable from the output of a model of the cochlea. The finite bandwidth of this model causes the second-formant peak to be influenced by higher formants. A possible articulatory correlate of F_2', suggested by Kuhn (1975), is that it may indicate the length of the mouth cavity (the cavity anterior to the point of maximum tongue constriction), at least for the more constricted vowels. Kuhn also suggests that emphasis on the mouth cavity as a perceptual variable may help in speaker normalization, since Fant (1966) had deduced from acoustic data that the difference in shape (as opposed to size) between male and female vocal tracts lies more in the pharynx than the mouth cavity, which is more linearly scaled between different-sized vocal tracts. The implication of all this is that the perceptual system uses much cruder information than is needed to specify completely an area function, to perceive a vowel category.

There is another reason why a heuristic approach to vowel perception might be advantageous. Stevens (1972) has shown that places of articulation for consonants and vowels occur at points where a small perturbation in articulation gives a minimum of perturbation in the acoustic output. If the perceptual system were capable of deriving an exact area function from the acoustic data, then this choice could be a disadvantage. However, if perception operates heuristically, the advantages of a sloppy articulation producing a relatively stable acoustic output, which is then mapped onto some idealized articulation, becomes obvious. Studies of articulation under abnormal conditions, such as might occur with a pipe held between the teeth (Lindblom, 1972), indicate that the articulation is drastically changed in order to attain a more nearly constant acoustic result. In addition, X-ray studies of vowel articulation (Ladefoged, DeClerk, Papçun, and Lindau, 1972) show that different speakers use very different tongue positions to produce the same phonetic vowel; again suggesting that some acoustic criterion must be satisfied. If, then, perception is mediated by articulatory variables, this articulation is unlikely to be equivalent to that of the speaker--as application of algorithms such as Atal's (1974) would lead to--or to that of the listener--as implied by motor theories. Rather, we must assume some more abstract form that is neither subject to their limitations nor capable of their variations.

Context and Prosodic Variables

Although this paper has concentrated on the problems of speech perception at the phonemic level, it would be misleading not to mention the further complications introduced by considering the perception of speech over segments longer than the isolated syllable or word.

Normal continuous speech varies from word to word in the precision with which it is articulated, and there appears to be an intimate relationship between the articulatory precision afforded a word and the ease with which that word could have been predicted by the listener. A word isolated from a predictable context is not as intelligible as the same word isolated from a less-predictable context (Lieberman, 1963). The complement of this observation is that listeners can use context to make up for poor articulation (Rubenstein and Pollack, 1963; Lieberman, 1967).

The ability to use context as an aid to perception is one virtue of an analysis-by-synthesis model of perception, but an impressive quantitative account of the interaction between context and stimulus information has been given within a passive framework by Morton (1970). In this model (see Figure 5), contextual constraints are seen as imposing a variable criterion on the decision mechanisms underlying word recognition, so that expected words are subject to a laxer criterion than are unexpected, and so they require less sensory information to produce a percept. Morton (see also Morton and Broadbent, 1967) contrasts this type of model with active models, which he maintains would predict a change in sensitivity rather than a change in criterion setting. This is presumably because the more expected word would be put up as a candidate to the analysis-by-synthesis comparator earlier than the less expected word, and so would find the stimulus trace in a less decayed or overwritten form.

But there is more to the perception of connected speech than the use of the context it supplies, for if words spoken in isolation are concatenated, the intelligibility of the resulting speech is very low (Stowe and Hampton, 1961). As Huggins (1972b) has observed, this is particularly striking when one considers that the intelligibility of the individual words that constitute this speech is higher than that of the same words spoken fluently. Presumably, then, perception of connected speech does not proceed solely as a sequence of serial decisions of phonemic or word size helped by phonological, syntactic, and semantic constraints, but is augmented by suprasegmental factors such as intonation and rhythm. Prosody obviously provides such information as where stress falls in a sentence, whether a question is being asked, what the mood of the speaker is, and so on (see review by Fry, 1970), but it perhaps also plays a more dynamic role in perception. It may serve to direct the listener's attention toward potentially informative parts of the speech stream (Cutler, 1975) and to segment the stream into chunks that are then candidates for higher-level units of analysis.

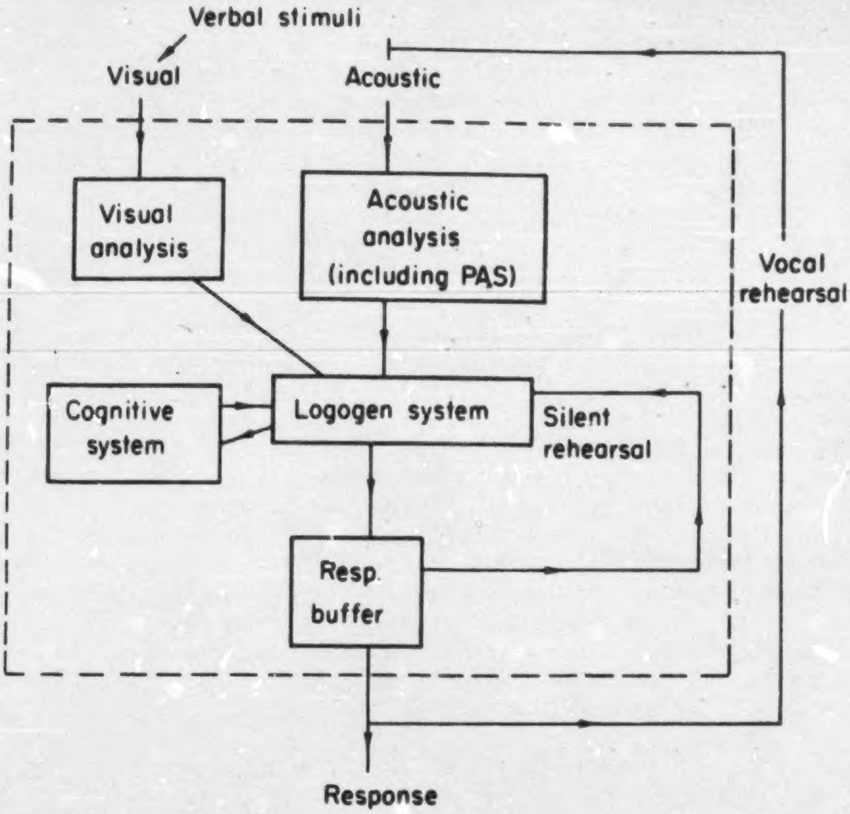
Verbal stimuli
Visual
Acoustic
Visual analysis
Acoustic analysis (including PAS)
Vocal rehearsal
Cognitive system
Logogen system
Silent rehearsal
Resp. buffer
Response

Figure 5: Morton's "Logogen Model." While this model makes no attempt to describe the auditory-to-phonemic stage in speech perception, it provides a useful summary of possible mechanisms before and after this stage. The persistence of a brief, overwritable precategorical acoustical store (PAS) is used to explain modality-specific recency effects in short-term serial recall, and is invoked in this paper to explain perceptual differences between different phonemic classes of speech sound. The logogen system provides an interface between sensory information and knowledge of the language and the world. The system consists of morpheme-sized units that can be biased by the cognitive system on the basis of its expectancies. This aspect of the model handles quantitatively changes in recognition accuracy of words with changing frequency, context, and signal-to-noise levels (Figure 1, Morton, 1970; reproduced with permission of publisher and author).

Experimental evidence on the importance of prosodic variables in perception is scattered, but what is available suggests that they have been unjustly neglected. Martin (1972) has discussed the theoretical implications of rhythm in perception, and Huggins (1972b) has provided a brief but useful review of the perceptual significance of prosody in connected speech. One of the points to emerge from this review is that listeners put more trust in prosodic information than they do in segmental when the two are in conflict. Wingfield and Klein (1971) examined this question by playing to subjects sentences whose intonation indicated that a major syntactic boundary was occurring at a point that was incompatible with the words in the sentence. Subjects' transcriptions of these mismatched sentences occasionally included word substitutions that changed the syntactic structure to be compatible with the intonation. Prosody seems also to be trusted more at the word level, since words read in a foreign accent tend to be heard as having the spoken, though incorrect, stress pattern even if this sacrifices useful segmental cues (Bansal, 1966; cited in Huggins, 1972b). This strategy is perhaps a wise one when we consider the relative resistance to distortion of prosodic and segmental information. Speech that is so severely band-limited that the overall intelligibility is only 30 percent still carries enough prosodic information for the stress pattern of the words to be correctly perceived (Kozhevnikov and Chistovich, 1965). The same is true for hummed speech (Svennson, 1974), which carries no segmental information. Using spectrally rotated speech, which again gives no segmental information, Blesser (1969) has found that for some sentences the syntactic structure of the transcribed sentence corresponds closely with that of the original.

While these experiments indicate that prosody furnishes useful information about stress patterns, and perhaps about syntactic structure, in the absence of segmental information, there has been little attempt to model how the interaction between syntactic information obtained lexically and that obtained prosodically is achieved. However, there are two approaches to sentence perception that might provide a suitable framework for approaching this interaction.

Bever (1970) has described a heuristic approach to the perception of sentences in which words are grouped together according to strategies based mainly on the grammatical class of the words. These strategies suggest how the major constituents of a sentence could be extracted, and provide a possible theoretical link between syntactic processing mechanisms and the use of prosodic information. If indeed prosody can help to determine syntactically useful segments, then strategies using this information could readily be incorporated into the type of scheme that Bever suggests. Experiments by Scholes (1971a, 1971b) provide a start on delimiting the usefulness of prosodic information for resolving syntactic ambiguity. In a similar vein, work done by Lea (1973) on the use of prosodic information as an aid to automatic speech recognition shows that the ends of major syntactic constituents (except noun-phrase/verbal boundaries) can be detected quite reliably by looking for a drop in the pitch contour.

Perhaps the most explicit model for the perception of sentences is the computer program by Winograd (1972) that allows typed communication in English about a world of colored blocks. The program uses a grammar, based on Halliday's (1967a, 1970) systemic grammar, that permits a rich interaction between syntactic and semantic constraints during the process of understanding an input sentence. Its dynamic use of semantic information provides a valuable constraint on the possible syntactic parsings of a sentence since semantically anomalous parsings can be rejected rapidly. For example, an attempt to parse "He gave the boy plants to water" as if it had the same structure as "He gave the house plants to charity" would be rejected when the semantic anomaly between boy and plants was detected. While such semantic constraints could also guide the perception of spoken language, there will also be additional constraints imposed by the prosodic information in the spoken sentence. Halliday's (1967b) own work on the relation between grammar and intonation might serve as a starting point for this link, particularly as Smith and Goodenough (1971) have found Halliday's analysis useful in explaining the time it takes subjects to answer questions asked in a variety of intonations following different introductory sentences.

CONCLUDING REMARKS

The experiments reviewed in the middle sections of this chapter have had some success at describing in information-processing terms possible mechanisms at early stages in the perception of brief, syllable-length segments of speech. Their success is to some extent a reflection of that of information-processing ideas in vision (e.g., Turvey, 1973), since the paradigms or the methodology have often been taken directly from visual work (e.g., Darwin, Turvey, and Crowder, 1972; Pisoni and Tash, 1974). However, the success of information-processing approaches in vision has been largely confined to "perception at a glance," a term that equally describes the scope of the speech work. We should ask whether our present enthusiastic pursuits with information-processing techniques are more likely to secure a golden fleece or a red herring. Are the tools at our disposal likely to lead to any real understanding of the true complexity of speech perception? What little we know about the perception of extended utterances and what little we know about the way in which the cues to phonological categories change when they occur in fluent speech both give pause to any simple attempt to relate auditory "perception at a glance" to the perception of fluent speech. Here, as elsewhere, the techniques of the experimental psychologist fall short of the task that faces him. Too often do psychological techniques that promise to illumine basic perceptual processes end up instead raising problems confined to their own methodology, which shed comparatively little light on the original problem.

What is needed is a way of modeling the speech perceptual process that is at once sufficiently complex to allow the richness of the system to be adequately represented and yet sufficiently transparent to provide insight. Computer programs for the synthesis of speech by rule already provide such a modeling medium for the phonologist (Mattingly, 1971), but interaction between speech psychologists and those concerned with automatic recognition of speech has until recently been minimal (Newell, 1971; Hyde, 1972), partly because of the tremendous technological problems of dealing with auditory signals. But as these problems are overcome, perhaps we can look forward to program-based models of the perceptual process being used to stimulate, and in turn be modified by, observations by psychologists on how the human brain perceives speech.

REFERENCES

Ades, A. E. (1973) Some effects of adaptation on speech perception. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 111, 121-129.

Ades, A. E. (1974a) How phonetic is selective adaptation? Experiments on syllable position and vowel environment. Percept. Psychophys. 16, 61-66.

Ades, A. E. (1974b) Bilateral component in speech perception? J. Acoust. Soc. Amer. 56, 610-616.

Ali, L, T. Gallagher, J. Goldstein, and R. Daniloff. (1971) Perception of coarticulated nasality. J. Acoust. Soc. Amer. 49, 538-540.

Allen, J. and M. P. Haggard. (1973) Dichotic backward masking of acoustically similar vowels. Speech Perception, Report on Speech Research in Progress (Psychology Department, The Queen's University, Belfast) Series 2, no. 3, 35-39.

Atal, B. S. (1974) Towards determining articulatory parameters from the speech wave. Paper presented at the International Congress of Acoustics, London.

Bailey, P. (1973) Perceptual adaptation for acoustical features in speech. Speech Perception, Report on Speech Research in Progress (Psychology Department, The Queen's University, Belfast) Series 2, no. 2, 29-34.

Bansal, R. K. (1966) The intelligibility of Indian English. Unpublished Ph.D. thesis, London University.

Bastian, J., P. D. Eimas, and A. M. Liberman. (1961) Identification and discrimination of a phonemic contrast induced by a silent interval. J. Acoust. Soc. Amer. 33, 842(A).

Bever, T. G. (1970) The cognitive basis for linguistic structures. In Cognition and the Development of Language, ed. by J. R. Hayes. (New York: Wiley).

Blakemore, C. and F. W. Campbell. (1969) On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. J. Physiol. 203, 237-260.

Blesser, B. (1969) Perception of spectrally rotated speech. Unpublished Ph.D. dissertation, Massachusetts Institute of Technology

Bregman, A. S. and J. Campbell. (1971) Primary auditory stream segregation and perception of order in rapid sequences of tones. J. Exp. Psychol. 89, 244-249.

Broadbent, D. E. (1958) Perception and Communication. (Oxford: Pergamon Press).

Broadbent, D. E. and P. Ladefoged. (1957) On the fusion of sounds reaching different sense organs. J. Acoust. Soc. Amer. 29, 708-710.

Campbell, F. W. and J. G. Robson. (1968) Application of Fourier analysis to the visibility of gratings. J. Physiol. 197, 551-566.

Carlson, R., G. Fant, and B. Granstrom. (1973) Two-formant models, pitch and vowel perception. Paper presented at the Symposium on Auditory Analysis and Perception of Speech, 21-24 August, Leningrad.

Carlson, R., B. Granstrom, and G. Fant. (1970) Some studies concerning perception of isolated vowels. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR-2/3, 19-35.

Chomsky, N. (1970) Phonology and reading. In Basic Studies on Reading, ed. by H. Levin and J. P. Williams. (New York: Basic Books).

Chomsky, N. and M. Halle. (1968) The Sound Pattern of English. (New York: Harper & Row).

Christie, W. M., Jr. (1974) Some cues for syllable juncture perception in English. J. Acoust. Soc. Amer. 55, 819-821.

Clowes, M. B. (1971) On seeing things. Artificial Intelligence 2, 79-116.

Cole, R. A. and B. Scott. (1974a) The phantom in the phoneme: Invariant cues for stop consonants. Percept. Psychophys. 15, 101-107.

Cole, R. A. and B. Scott. (1974b) Toward a theory of speech perception. Psychol. Rev. 81, 348-374.

Cooper, F. S., P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman. (1952) Some experiments on the perception of synthetic speech sounds. J. Acoust. Soc. Amer. 24, 597-606.

Cooper, W. (1974) Adaptation of phonetic feature analyzers for place of articulation. J. Acoust. Soc. Amer. 56, 617-627.

Cooper, W. (1975) Selective adaptation to speech. In Cognitive Theory, ed. by E. Restle, R. M. Shiffrin, J. N. Castellan, H. Lindman, and D. B. Pisoni. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Cooper, W. E. and S. Blumstein. (1974) A "labial" feature analyzer in speech perception. Percept. Psychophys. 15, 591-600.

Corteen, R. S. and B. Wood. (1972) Autonomic responses to shock-associated words in an unattended channel. J. Exp. Psychol. 94, 308-313.

Crowder, R. G. (1971) The sound of vowels and consonants in immediate memory. J. Verbal Learn. Verbal Behav. 10, 587-596.

Crowder, R. G. (1973) Representation of speech sounds in precategorical acoustic storage. J. Exp. Psychol. 98, 14-24.

Crowder, R. G. and J. Morton. (1969) Precategorical acoustic storage (PAS). Percept. Psychophys. 5, 365-373.

Cudahy, E. and B. Leshowitz. (1974) Effects of a contralateral interference tone on auditory recognition. Percept. Psychophys. 15, 16-20.

Cutler, A. (1975) Rhythmic factors in the determination of perceived stress. Paper presented at the 89th Meeting of the Acoustical Society of America, 8-11 April, Austin, Tex.

Cutting, J. E. (1972) Ear advantage for stops and liquids in initial and final position. Haskins Laboratories Status Report on Speech Research SR-31/32, 57-65.

Daniloff, R. and K. Moll. (1968) Coarticulation of lip-rounding. J. Speech Hearing Res. 11, 707-721.

Darwin, C. J. (1969) Auditory perception and cerebral dominance. Unpublished Ph.D. thesis, University of Cambridge.

Darwin, C. J. (1971a) Ear differences in the recall of fricatives and vowels. Quart. J. Exp. Psychol. 23, 46-62.

Darwin, C. J. (1971b) Dichotic backward masking of complex sounds. Quart. J. Exp. Psychol. 23, 386-392.

Darwin, C. J. (1973) Ear differences and hemispheric specialization. In The Neurosciences, Third Study Program, ed. by F. O. Schmitt and F. G. Worden. (Cambridge, Mass.: MIT Press), pp. 57-63.

Darwin, C. J. and A. D. Baddeley. (1974) Acoustic memory and the perception of speech. Cog. Psychol. 6, 41-60.

Darwin, C. J. and S. A. Brady. (1975) Voicing and juncture in stop-lateral clusters. Poster presented at the 89th Meeting of the Acoustical Society of America, 8-11 April, Austin, Tex.

Darwin, C. J., M. T. Turvey, and R. G. Crowder. (1972) An auditory analogue of the Sperling partial report procedure: Evidence for brief auditory storage. Cog. Psychol. 3, 255-267.

Delattre, P. C., A. M. Liberman, and F. S. Cooper. (1955) Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Amer. 27, 769-773.

Denes, P. (1955) Effect of duration on perception of voicing. J. Acoust. Soc. Amer. 27, 761-764.

Dixit, R. P. and P. F. MacNeilage. (1972) Coarticulation of nasality: Evidence from Hindi. J. Acoust. Soc. Amer. 52, 131(A).

Eimas, P. D., W. E. Cooper, and J. D. Corbit. (1973) Some properties of linguistic feature detectors. Percept. Psychophys. 13, 247-252.

Eimas, P. D. and J. D. Corbit. (1973) Selective adaptation of linguistic feature detectors. Cog. Psychol. 4, 99-109.

Evans, E. F. and I. C. Whitfield. (1964) Classification of unit repsonses in the auditory cortex of the unanesthetized and unrestrained cat. J. Physiol. (London) 171, 476-493.

Fant, C. G. M. (1960) Acoustic Theory of Speech Production. (The Hague: Mouton).

Fant, C. G. M. (1964) Auditory patterns of speech. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR-3, 16-20.

Fant, G. (1966) A note on vocal tract size factors and nonuniform F-pattern scalings. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR-4, 22-30.

Fant, G. (1967) Auditory patterns of speech. In Models for the Perception of Speech and Visual Form, ed. by W. Wathen-Dunn. (Cambridge, Mass.: MIT Press).

Fant, G., J. Liljencrants, V. Malac, and B. Borovičkova. (1970) Perceptual evaluation of coarticulation effects. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR-1, 10-13.

Fletcher, H. (1929) Speech and Hearing. (New York: Van Nostrand).

Foss, D. J. and D. A. Swinney. (1973) On the psychological reality of the phoneme: Perception, identifictiton and consciousness. J. Verbal Learn. Verbal Behav. 12, 246-257.

Fourcin, A. J. (1968) Speech source inference. IEEE Trans. Audio Electroacoust. AU-16, 65-67.

Fry, D. B. (1970) Prosodic phenomena. In Manual of Phonetics, ed. by B. Malmberg. (Amsterdam: North-Holland).

Fry, D. B., A. S. Abramson, P. D. Eimas, and A. M. Liberman. (1962) The identification and discrimination of synthetic vowels. Lang. Speech 5, 171-189.

Fujisaki, H. and T. Kawashima. (1968) The influence of various factors on the identification and discrimination of synthetic speech sounds. Paper presented at the 6th International Congress on Acoustics, August, Tokyo, Japan.

Gårding, E. (1967) Internal Juncture in Swedish. Travaux de L'Institut de Phonétique de Lund VI. (Lund: Gleerup).

Gazzaniga, M. S. and R. W. Sperry. (1967) Language after section of the cerebral commissures. Brain 90, 131-148.

Godfrey, J. J. (1974) Perceptual difficulty and the right-ear advantage for vowels. Brain Lang. 1, 323-336.

Gombrich, E. (1960) Art and Illusion. (Princeton, N. J.: Princeton University Press).

Gough, P. B. (1972) One second of reading. In Langauge by Ear and by Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Guzman, A. (1968) Computer recognition of three-dimensional objects in a visual scene. MAC Technical Report (Project MAC, MIT) 59.

Hadding-Koch, K. and M. Studdert-Kennedy. (1964) An experimental study of some intonation contours. Phonetica 11, 175-185.

Haggard, M. P. (1971) Theoretical issues in speech perception. Speech Synthesis and Perception (Psychological Laboratory, University of Cambridge) 4, 1-16.

Haggard, M. P., J. M. Corrigall, and A. E. Legg. (1971) Perceptual factors in articulatory defects. Folia Phoniat. 23, 33-40.

Halliday, M. A. K. (1967a) Notes on transitivity and theme in English. J. Ling. 3, 37-81 and 4, 179-215.

Halliday, M. A. K. (1967b) Intonation and Grammar in British English. (The Hague: Mouton).

Halliday, M. A. K. (1970) Functional diversity in language as seen from a consideration of modality and mood in English. Foundations of Language 6, 322-361.

Harris, K. S. (1958) Cues for the discrimination of American English fricatives in spoken syllables. Lang. Speech 1, 1-17.

Holmes, J. N., I. G. Mattingly, and J. N. Shearme. (1964) Speech synthesis by rule. Lang. Speech 7, 127-143.

Hubel, D. H. and T. N. Wiesel. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. (London) 160, 106-154.

Huggins, A. W. F. (1972a) Just noticeable differences for segment duration in natural speech. J. Acoust. Soc. Amer. 51, 1270-1278.

Huggins, A. W. F. (1972b) On the perception of temporal phenomena in speech. J. Acoust. Soc. Amer. 51, 1279-1290.

Hyde, S. R. (1972) Automatic speech recognition: An initial survey of the literature. In Human Communication: A Unified View, ed. by E. E. David and P. B. Denes. (New York: McGraw-Hill).

Johnston, J. C. and J. L. McClelland. (1974) Perception of letters in words: Seek not and ye shall find. Science 184, 1192-1194.

Kay, R. H. and D. R. Matthews. (1972) On the existence in human auditory pathways of channels selectively tuned to the modulation present in frequency-modulated tones. J. Physiol. (London) 225, 657-677.

Kirstein, E. F. (1970) Selective listening for temporally staggered dichotic CV syllables. J. Acoust. Soc. Amer. 48, 95(A).

Klatt, D. J. (1973) Voice-onset time, frication, and aspiration in word-initial consonant clusters. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 109, 124-136.

Kozhvnikov, V. A. and L. A. Chistovich. (1965) Speech: Articulation and Perception. Translated from Russian. (Washington, D.C.: U. S. Department of Commerce, Clearinghouse for Federal Scientific and Technical Information).

Kuhn, G. M. (1973) A two-pass procedure for synthesis-by-rule. J. Acoust. Soc. Amer. 54, 339(A).

Kuhn, G. M. (1975) On the front cavity resonance, and its possible role in speech perception. Haskins Laboratories Status Report on Speech Research SR-41, 105-116.

Lackner, J. R. and L. M. Goldstein. (1974) The psychological representation of speech sounds. Cognition 2, 279-298.

Ladefoged, P. (1966) The nature of general phonetic theories. Languages and Linguistics (Georgetown University), Monograph No. 18, 27-42.

Ladefoged, P. and D. E. Broadbent. (1957) Information conveyed by vowels. J. Acoust. Soc. Amer. 29, 98-104.

Ladefoged, P., J. L. DeClerk, G. Papçun, and M. Lindau. (1972) An auditory-motor theory of speech production. Working Papers in Phonetics (Linguistics Department, University of California at Los Angeles) 22, 48-75.

Lea, W. A. (1973) An approach to syntactic recognition without phonemics. IEEE Trans. Audio Electroacoust. AU-21, 249-258.

Lehiste, I. (1960) An acoustic-phonetic study of internal open juncture. Phonetica, Suppl. 5.

Lehiste, I. and L. Shockey. (1972) On the perception of coarticulation effects in English VCV syllables. Working Papers in Linguistics (Linguistics Department, Ohio State University) 12, 78-86.

Lenneberg, E. H. (1962) Understanding language without ability to speak: A case report. *J. Abnormal Social Psychol. 65*, 419-425.

Liberman, A. M., F. S. Cooper, K. S. Harris, and P. F. MacNeilage. (1962) A motor theory of speech perception. In *Proceedings of the Speech Communication Seminars*, vol. 2. (Stockholm: Royal Institute of Technology).

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. *Psychol. Rev. 74*, 431-461.

Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1952) The role of selected stimulus variables in the perception of the unvoiced-stop consonants. *Amer. J. Psychol. 65*, 497-516.

Liberman, A. M., P. Delattre, and F. S. Cooper. (1958) Some cues for the distinction between voiced and voiceless stops in initial position. *Lang. Speech 1*, 153-167.

Liberman, A. M., P. C. Delattre, F. S. Cooper, and L. J. Gerstman. (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychol. Monogr. 68*, 8, Whole No. 379.

Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol. 54*, 358-368.

Liberman, A. M., F. Ingemann, L. Lisker, P. C. Delattre, and F. S. Cooper. (1959) Minimal rules for synthesizing speech. *J. Acoust. Soc. Amer. 31*, 1490-1499.

Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In *Coding Processes in Human Memory*, ed. by A. W. Melton and E. Martin. (New York: Wiley), pp. 307-334.

Licklider, J. C. R. (1951) A duplex theory of pitch perception. *Experientia 7*, 128-134.

Lieberman, P. (1963) Some effects of semantic and grammatical context on the production and perception of speech. *Lang. Speech 6*, 172-187.

Lieberman, P. (1965) On the acoustic basis of the perception of intonation by linguists. *Word 21*, 40-54.

Lieberman, P. (1967) Intonation, perception, and language. *Research Monograph*, No. 38. (Cambridge, Mass.: MIT Press).

Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. *J. Acoust. Soc. Amer. 35*, 1773-1781.

Lindblom, B. (1972) Phonetics and the description of language. In *Proceedings of the 7th International Congress of Phonetic Sciences*. (The Hague: Mouton), pp. 63-93.

Lindblom, B. E. F. and M. Studdert-Kennedy. (1967) On the role of formant transitions in vowel recognition. *J. Acoust. Soc. Amer. 42*, 830-843.

Lisker, L. (1957a) Closure duration and the intervocalic voiced-voiceless distinction in English. *Language 33*, 42-49.

Lisker, L. (1957b) Minimal cues for separating /w,r,l,y/ in intervocalic position. *Word 13*, 256-267.

Lisker, L. (1961) Voicing lag in clusters of stop plus /r/. *Speech Research and Instrumentation*, Ninth Final Report, Haskins Laboratories, Appendix A-II.

Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. *Word 20*, 384-422.

Lisker, L. and A. S. Abramson. (1967) Some effects of context on voice onset time in English stops. *Lang. Speech 10*, 1-28.

Lisker, L. and A. S. Abramson. (1970) The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th International Congress of Phonetic Sciences, 1967*. (Prague: Academia), pp. 563-567.

Lukatela, G. (1973) Pitch determination by adaptive autocorrelation method. *Haskins Laboratories Status Report on Speech Research* *SR-33*, 185-193.

Mackworth, A. K. (1973) Interpreting pictures of polyhedral scenes. Paper presented at the 3rd International Joint Conference on Artificial Intelligence.

MacNeilage, P. F. and P. Ladefoged. (in press) The production of speech. In *Handbook of Perception*, vol. 7, ed. by E. C. Carterette and M. P. Friedman. (New York: Academic Press).

Malecot, A. (1956) Acoustic cues for nasal consonants. *Language* *32*, 274-284.

Malmberg, B. (1955) The phonetic basis for syllable division. *Studia Linguistica* *9*, 80-87.

Martin, J. G. (1972) Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychol. Rev.* *79*, 487-509.

Massaro, D. M. (1970) Preperceptual auditory images. *J. Exp. Psychol.* *85*, 411-417.

Mattingly, I. G. (1968) Synthesis by rule of General American English. Ph.D. dissertation, Yale University. (Issued as Supplement to *Haskins Laboratories Status Report on Speech Research*.)

Mattingly, I. G. (1971) Synthesis by rule as a tool for phonological research. *Lang. Speech* *14*, 47-56.

McNeill, D. and L. Lindig. (1973) The perceptual reality of phonemes, syllables, words, and sentences. *J. Verbal Learn. Verbal Behav.* *12*, 417-430.

Moll, K. L. and R. G. Daniloff. (1971) An investigation of the timing of velar movements during speech. *J. Acoust. Soc. Amer.* *50*, 678-684.

Morton, J. (1970) A functional model for memory. In *Models of Human Memory*, ed. by D. A. Norman. (New York: Academic Press).

Morton, J. and D. E. Broadbent. (1967) Passive versus active recognition models, or is your homunculus really necessary? In *Models for the Perception of Speech and Visual Form*, ed. by W. Wathen-Dunn. (Cambridge, Mass.: MIT Press).

Murrell, G. A. and J. Morton. (1974) Word recognition and morphemic structure. *J. Exp. Psychol.* *102*, 963-968.

Neisser, U. (1967) *Cognitive Psychology*. (New York: Appleton-Century-Crofts).

Newell, A. (1971) *Speech-Understanding Systems: Final Report of a Study Group* (Computer Science Department, Carnegie-Mellon University).

O'Connor, J. D., L. H. Gerstman, A. M. Liberman, P. C. Delattre, and F. S. Cooper. (1957) Acoustic cues for the perception of initial /w,j,r,l/ in English. *Word* *13*, 24-43.

Öhman, S. E. G. (1966) Coarticulation in VCV utterances. *J. Acoust. Soc. Amer.* *39*, 151-168.

Öhman, S. E. G. (1967) Numerical model of coarticulation. *J. Acoust. Soc. Amer.* *41*, 310-320.

Peterson, G. E. and H. L. Barney. (1952) Control methods used in a study of the identification of vowels. *J. Acoust. Soc. Amer.* *24*, 175-184.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Ph.D. thesis, University of Michigan. (Issued as Supplement to *Haskins Laboratories Status Report on Speech Research*.)

Pisoni, D. B. (1972) Perceptual processing time for consonants and vowels. *Haskins Laboratories Status Report on Speech Research* *SR-31/32*, 83-92.

Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Percept. Psychophys.* *13*, 253-260.

Pisoni, D. B. (1975) Dichotic listening and processing phonetic features. In *Cognitive Theory*, vol. 1, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, and H. Lindman. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Pisoni, D. B. and S. D. McNabb. (1974) Dichotic interactions and phonetic feature processing. *Brain Lang. 1*, 351-362.

Pisoni, D. B. and J. Tash. (1974) Reaction times to comparisons within and across phonetic categories. *Percept. Psychophys. 15*, 285-290.

Porter, R. J. (1971) The effect of temporal overlap on the perception of dichotically and monotically presented CV syllables. *J. Acoust. Soc. Amer. 50*, 129(A).

Porter, R. J., D. P. Shankweiler, and A. M. Liberman. (1969) Differential effects of binarual time differences on perception of stop consonants and vowels. Paper presented at the 77th Meeting of the American Psychological Association, Washington, D.C.

Posner, M. I. and R. F. Mitchell. (1967) Chronometric analysis of classification. *Psychol. Rev. 74*, 392-409.

Rand, T. C. (1971) Vocal tract size normalization in the perception of stop consonants. *Haskins Laboratories Status Report on Speech Research SR-25/26*, 141-146.

Rand, T. C. (1974) Dichotic release from masking for speech. *J. Acoust. Soc. Amer. 55*, 678-680.

Raphael, L. J. (1972) Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *J. Acoust. Soc. Amer. 51*, 1296-1303.

Reicher, G. M. (1969) Perceptual recognition as a function of meaningfulness of stimulus material. *J. Exp. Psychol. 81*, 275-280.

Rubin, P., M. T. Turvey, and P. Van Gelder. (in press) Semantic influences on phonological processing. *Haskins Laboratories Status Report on Speech Research SR-44*.

Rubinstein, H. and I. Pollack. (1963) Word predictability and intelligibility. *J. Verbal Learn. Verbal Behav. 2*, 147-158.

Savin, H. and T. G. Bever. (1970) The nonperceptual reality of the phoneme. *J. Verbal Learn. Verbal Behav. 3*, 295-302.

Sawusch, J. R., D. B. Pisoni, and J. E. Cutting. (1974) Category boundaries for linguistic and nonlinguistic dimensions of the same stimuli. Paper presented at the 87th Meeting of the Acoustical Society of America, April, New York.

Schatz, C. (1954) The role of context in the perception of stops. *Language 30*, 47-56.

Scholes, R. J. (1971a) *Acoustic Cues for Constituent Structure*. (The Hague: Mouton).

Scholes, R. J. (1971b) On the spoken disambiguation of superficially ambiguous sentences. *Lang. Speech 14*, 1-11.

Schwartz, M. F. (1967) Transitions in American English /s/ as cues to the identity of adjacent stop consonants. *J. Acoust. Soc. Amer. 42*, 897-899.

Schwartz, M. F. (1968) Identification of speaker sex from isolated voiceless fricatives. *J. Acoust. Soc. Amer. 43*, 1178-1179.

Selfridge, O. G. and U. Neisser. (1960) Pattern recognition by machine. *Sci. Amer. 203* (Aug.), 60-68.

Shankweiler, D. P., W. Strange, and R. Verbrugge. (in press) Speech and the problem of perceptual constancy. In *Perceiving, Acting and Comprehending: Toward an Ecological Psychology*, ed. by R. Shaw and J. Bransford. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.). [Also in Haskins Laboratories *Status Report on Speech Research* (this issue).]

Shankweiler, D. P. and M. Studdert-Kennedy. (1967) Identification of consonants and vowels presented to left and right ears. *Quart. J. Exp. Psychol. 19*, 59-63.

Shearme, J. N. and J. N. Holmes. (1962) An experimental study of the classification of sounds in continuous speech according to their distribution in the F1-F2 plane. In Proceedings of the 4th International Congress of Phonetic Science, Helsinki, 1961. (The Hague: Mouton).

Small, A. M. (1970) Periodicity pitch. In Foundations of Modern Auditory Theory, vol. 1, ed. by J. V. Tobias. (New York: Academic Press).

Smith, F. and C. Goodenough. (1971) Effects of context, intonation, and voice on the reaction time to sentences. Lang. Speech 14, 241-250.

Stevens, K. N. (1972) The quantal nature of speech: Evidence from articulatory-acoustic data. In Human Communication: A Unified View, ed. by E. E. David and P. B. Denes. (New York: McGraw-Hill).

Stevens, K. and A. S. House. (1972) Speech perception. In Foundations of Modern Auditory Theory, vol. 2, ed. by J. V. Tobias. (New York: Academic Press).

Stevens, K. N. and D. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Amer. 55, 653-659.

Stowe, A. N. and D. B. Hampton. (1961) Speech synthesis with pre-recorded syllables and words. J. Acoust. Soc. Amer. 33, 810-811.

Strevens, P. (1960) Spectra of fricative noise. Lang. Speech 3, 32-49.

Studdert-Kennedy, M. (1974) The perception of speech. In Current Trends in Linguistics, vol. 12: Phonetics, ed. by T. A. Sebeok. (The Hague: Mouton).

Studdert-Kennedy, M. and F. S. Cooper. (1966) High-performance reading machines for the blind. In Proceedings of the International Conference on Sensory Devices for the Blind. (London: St. Dunstan's), pp. 317-340.

Studdert-Kennedy, M., D. Shankweiler, and S. Schulman. (1970) Opposed effects of a delayed channel on perception of dichotically and monotically presented CV syllables. J. Acoust. Soc. Amer. 48, 599-602.

Summerfield, A. Q. (1974) Toward a detailed model for the perception of voicing contrasts. Speech Perception, Report on Speech Research in Progress (Psychology Department, The Queen's University, Belfast) Series 2, no. 3, 1-26.

Summerfield, A. Q. and M. P. Haggard, (1974) Perceptual processing of multiple cues and contexts: Effects of following vowel upon stop consonant voicing. J. Phonetics 2, 279-295.

Svennson, S-G. (1974) Prosody and grammar in speech perception. MILUS, No. 2. (Institute of Linguistics, University of Stockholm).

Tartter, V. C. (1975) Selective adaptation of acoustic and phonetic detectors. Unpublished M.A. thesis, Brown University.

Tash, J. (1974) Selective adaptation of auditory feature detectors in speech perception. M.A. dissertation, University of Indiana. [Published in Research on Speech Perception (Department of Psychology, Indiana University), Progress Report No. 1.]

Treisman, A. M. (1960) Contextual cues in selective listening. Quart. J. Exp. Psychol. 12, 242-248.

Treisman, A. M. and J. G. A. Riley. (1969) Is selective attention selective perception or selective response? A further test. J. Exp. Psychol. 79, 27-34.

Turvey, M. T. (1973) On peripheral and central processes in vision. Psychol. Rev. 80, 1-52.

Wang, W. S-Y. and C. Fillmore. (1961) Intrinsic cues and consonant perception. J. Speech Hearing Res. 4, 130-136.

Warren, R. M. (1968) Verbal transformation effect and auditory perceptual mechanisms. Psychol. Bull. 70, 261-270.

Warren, R. M. and J. M. Ackroft. (1974) Dichotic verbal transformation: Evidence of separate neural processes for identical stimuli. J. Acoust. Soc. Amer., Suppl. 56, 54(A).

Wheeler, D. D. (1970) Processes in word recognition. Cog. Psychol. 1, 59-85.

Wightman, F. L. (1973) Pattern-transformation model of pitch. J. Acoust. Soc. Amer. 54, 407-416.

Wilson, J. P. (1973) Psychoacoustical and neurophysiological aspects of auditory pattern recognition. In The Neurosciences: Third Study Program, ed. by F. O. Schmitt and F. G. Worden. (Cambridge, Mass.: MIT Press).

Wingfield, A. and J. F. Klein. (1971) Syntactic structure and acoustic pattern in speech perception. Percept. Psychophys. 9, 23-25.

Winitz, H., M. E. Scheib, and J. A. Reeds. (1972) Identification of stops and vowels from the burst portion of /p,t,k/ isolated from conversational speech. J. Acoust. Soc. Amer. 51, 1309.

Winograd, T. (1972) Understanding Natural Language. (New York: Academic Press).

On the Dynamic Use of Prosody in Speech Perception*

C. J. Darwin[+]

ABSTRACT

Two roles that prosodic variables might play in the perception of speech are reviewed and two experiments described. Prosody is seen on the one hand as helping to direct attention to a particular speaker and to the potentially most informative parts of his speech. It is also seen as modifying the hypotheses a listener might entertain about a sentence as he listens to it. The use of models of perception taken from natural language understanding programs is suggested as a way to begin modeling this second role of prosody

There is little doubt that prosodic variables make a significant contribution to the intelligibility of speech. We lack, however, a suitable framework for modeling, in perception, the relation between prosodic variables and segmental information. Part of the reason for this is that many very different types of information can be conveyed by prosodic variables, and it would be a brave soul who denied any of them a role in intelligibility. The segmental distinctions carried by normally prosodic dimensions, such as change in pitch as a weak cue to voicing (Fujimura, 1961; Haggard, Ambler, and Callow, 1970) and lexical distinctions carried by stress (e.g., see Fry, 1970), are not particularly difficult to integrate into a scheme for the perception of lexical items from segmental cues. But these aspects of prosody exclude perhaps the bulk of the contribution that prosody makes to the intelligibility of speech. Sentences made up of concatenated words spoken in isolation are less intelligible than those spoken fluently (Stowe and Hampton, 1961; Abrams and Bever, 1969), a finding that is almost paradoxical when one considers (Huggins, 1972) that words excised from fluent speech are less intelligible than the same words spoken in isolation (Lieberman, 1963). The resolution of this paradox can be achieved in part by

---

*This paper was prepared for the Symposium on Dynamic Aspects of Speech Perception, held at the Instituut for Perceptie Onderzoek, Eindhoven, 4-6 August 1975. It is to be published in Structure and Process in Speech Perception, ed. by A. Cohen and S. G. Nooteboom. (Berlin: Springer-Verlag).

[+] Also University of Connecticut, Storrs; on leave of absence from University of Sussex, Brighton, England.

Acknowledgment: The first experiment reported here was carried out at the University of Sussex; the second, at the University of Connecticut. The work was supported in part by a grant from the British S.R.C.

[HASKINS LABORATORIES: Status Report on Speech Research SR-42/43 (1975)]

knowing that coarticulation is no respecter of word boundaries, but it is likely, if not compelling, that the prosodic information present in fluent speech helps perception of the fluent utterance. When coarticulatory artifacts are made less likely, prosodic perturbations less drastic than word splicing still influence intelligibility. When sentences sharing a common five-word portion, but having major syntactic boundaries at different places within this common string, are cross-spliced so that an intonation contour inappropriate to the sentence's syntactic structure is heard, the intelligibility of these mosaic sentences is lower than those with normal intonation (Wingfield and Klein, 1971; Wingfield, 1975). This technique of cross-splicing introduces two perturbations into the prosody; first it produces an abrupt change in both pitch and rhythm at the splice points, and second it gives an inappropriate placement of the intonational cues to the syntactic boundary. While both of these changes may contribute to the reduced intelligibility of the mosaic sentences, the latter certainly has a specific effect since subjects' transcriptions of these sentences include ones with the syntactic boundary at the point suggested by the intonation contour. This result can be compared with the outcome of other experiments that have shown that subjects can extract considerable information about the stress pattern of speech even when there is virtually no segmental information present, either by virtue of its being hummed (Svensson, 1974) or spectrally rotated (Blesser, 1969). If indeed information that is a potentially useful indicant of syntactic structure can be extracted independently of segmental information, do we need to postulate any interaction between segmental and prosodic information while the sentence is being perceived? Perhaps Wingfield and Klein's (1971) subjects simply reinterpreted the sentences they heard in the light of an independently perceived intonation contour. Not very dynamic, but certainly easy to model!

Life would be more interesting and speech perhaps easier to perceive if prosody played a more dynamic role. This paper aims to review the evidence that bears on a potentially dynamic role of prosody in the perception of speech and to present two new fragments of data that contribute to this discussion. The review will be based on two putative roles of prosodic variables: the role of rhythm and pitch contour in both allowing temporal prediction and giving continuity to an attended channel, and the role of prosodic variables in delimiting higher syntactic structures.

TEMPORAL PATTERNING, PREDICTION, AND ATTENTION

Reaction-time techniques provide a useful way of studying the ongoing process of sentence perception. The most extensively used technique, the phoneme-monitoring task, was introduced by Foss and Lynch (1969). Here, subjects, while listening to a sentence that they subsequently must recall, have to press a key whenever they hear a word beginning with a particular phoneme (usually /b/). Changes in this reaction time have been found to depend on a number of syntactic and semantic variables. Reaction time is longer when the target word occurs in a self-embedded than in a right-branching sentence (Foss and Lynch, 1969), in sentences with deleted rather than intact relative pronouns (Hakes and Cairns, 1970; Hakes and Foss, 1970), in sentences with reduced rather than intact complements (Hakes, 1972), and after high rather than low frequency words (Foss, 1969; at least for adjectives of different frequencies, Cairns and Foss, 1971). These differences have been interpreted as reflecting changes in the processing load imposed by the sentence (cf. Aaronson, 1968) and should not be thought of as reflecting directly the extraction of segmental information. The conscious access to the phonemic level that this task requires is made after (or at least is influenced by) decisions about what the word containing the target is. Reaction times are shorter if the target starts a word rather than a nonword (Rubin, Turvey, and Van Gelder, in press; see also Foss and Swinney, 1973; Treisman and Tuxworth, 1974, for discussion of related points).

The phoneme monitoring task has been used in two contexts that relate to the use of prosodic information. An experiment by Cutler and Foss (1973) provides a convenient starting point. Following an earlier finding that reaction time to an intital stop consonant was faster when it appeared at the beginning of a content than of a function word, they showed that this difference was attributable to differences in the stress with which function and content words are normally spoken. When stress and lexical category were varied independently, the faster reaction time accompanied the stressed word. Although unstressed function words gave slower reaction times than unstressed content words, they felt this could be owing to different degrees of stress.

This experiment alone can support a number of interpretations: (1) stressed words may be articulated more precisely so that there is more segmental information available, (2) stressed words may be processed more efficiently on account of their intrinsic stress, (3) stressed words may be processed more efficiently through the preceding stress pattern, suggesting that they will be stressed. This last hypothesis is compatible with recent comments on the use of rhythm by Martin (1972).

The first two interpretations attribute the faster reaction time for stressed words to factors intrinsic to the word itself, and can now be dismissed with some confidence. Shields, McHugh, and Martin (1974) measured reaction time to nonsense disyllables beginning with a target phoneme. The disyllable was pronounced with the stress either on the first or the second syllable. When the word occurred as part of a fluent sentence, subjects were faster to a target in a stressed syllable than in an unstressed one, provided that the target did not occur too close to the end of the sentence. However, when the same target words were spliced out and presented in isolated list form, the differences between stressed and unstressed syllables disappeared. The curious lack of difference when the targets occur at the end of the sentence may be simply a floor effect, since reaction times decreased throughout the sentence for both types of syllables. This study is complemented by a recent experiment by Cutler (cited in Cutler, 1975) in which target words were spliced into a sentence context, whose prosody suggested that the target word would or would not be stressed. She found that reaction time was faster when subjects expected a stressed word than when they did not. The stress difference thus seems to be independent of the intrinsic stress of a word and depends rather on the preceding prosodic pattern, which allows the subject to anticipate the forthcoming stressed word.

If some form of anticipation is the cause of the stressed-syllable advantage, then we might expect two further effects; some time should be necessary for subjects to get the rhythm of the utterance on which prosodic predictions might be based, and local disturbance of the rhythm should upset the stressed-syllable advantage. There is evidence for both. Aaronson (1968) asked her subjects to monitor a list of digits, spoken at a constant rate, for the occurrence of a target digit. There was a decrease of about 100-msec reaction time to the target over the first three serial positions. This occurred whether subjects subsequently had to recall the list of digits or not. Reaction times remained steady in subsequent serial positions for subjects who had only to monitor the lists, but they increased after about the third item until the end of the list for subjects who had to recall the lists as well as monitor. We will return to this latter finding below; for the present it is enough to note that the initial decrease in reaction time may well reflect subjects' accommodating to the rhythm of the list. Cutler (1975) showed that local temporal disturbances influence phoneme-monitoring reaction times. She found that inserting a quarter-second period of silence before a pair of monosyllabic words embedded in a fluently spoken sentence influenced the reaction time to a phoneme target at the beginning of either of the words. The reaction time was slowed to the second word, which was initially spoken with stress, and quickened to the unstressed first word. Cutler failed to find any advantage for the stressed word over the unstressed when the silent interval was absent; this may possibly be attributable to the target's position in the sentence (cf. Shields et al., 1974).

The picture emerging from these studies, then,is of subjects' attention being allocated preferentially toward the stressed syllables in a sentence (Shields et al., 1974). This is made easier by the timing of speech (at least in a stress-timed language) being determined by stressed syllables. Allen (1972), for example, showed that subjects can tap with less variability to stressed syllables than to unstressed (even when the sentence in which they were embedded was heard repeatedly), and Huggins (1972) finds subjects to be more tolerant of timing distortions that preserve stressed-syllable separations. As Cutler and Foss (1973) point out, allowing processing to be directed toward the stressed parts of the sentence allows the focus of the speaker's sentence to control the listener's perception. However, we might note in passing that it is not yet clear how much the results of these experiments depend on using a temporal response measure. The task of targeting for a phoneme might cause an undue advantage to those portions of the speech stream that are temporally predictable. It would strengthen the case if other measures, such as perhaps the detectability of feature substitutions, showed a similar increase in stressed syllables. Some support comes from an experiment by Dooling (1974). He found that subjects' immediate recall of sentences heard in noise was improved if they had previously been exposed to sentences with a similar stress pattern. His results also showed that repetition of stress pattern was much more effective at improving performance than was repetition of syntactic structure, when these two variables were independently manipulated.

Although Martin's (1972) theorizing emphasizes the predictive nature of rhythm, others have shown that pitch contour can play a similar role. The phenomenon of "primary auditory stream segregation" (Bregman and Campbell, 1971) illustrates this, and again, like Martin's work, embraces both speech and music. A random sequence of six notes (three high, three low), when played rapidly, will perceptually segment into a high and a low tune, despite lack of any greater rhythmic cohesion within than between tunes. The analogy with speech perhaps lies both in the use of frequency continuity of formants to help in their tracking (cf. Dorman, Cutting, and Raphael, 1975) and in the use of continuity of pitch to help in attending to one voice against competing sounds. This last point has been pursued in an experiment carried out in 1973 by Darwin and Davina Simmonds. While this experiment was motivated by the Bregman and Campbell (1971) findings, it does not in fact distinguish between rhythm and pitch, but looks at the influence of prosodic variables in general on the ability to attend to a particular speech source.

EXPERIMENT I

The technique of shadowing, although complex for the subject, provides a valuable way of studying the process of selective attention to a particular speaker. Treisman (1960) asked subjects to shadow a passage of continuous speech led to one ear and to ignore a similar passage led to the other. At some point during the passage the two channels were switched so that the passage that had been shadowed continued in the other ear. She found that subjects occasionally gave, as part of their shadowing response, words that had in fact occurred immediately after the switch on the ear they had been instructed to ignore, and that these intrusions were more common the more redundant the passages. Inasmuch as this effect has any sensory basis, we can ask whether the tendency for subjects to shadow words from the unattended ear is being determined by some semantic/syntactic priming, as Treisman maintained, or rather whether the continuity in rhythm and pitch across the ears at the switch point is sufficient to cause a momentary change in the ear from which the attended auditory input is drawn.

Our experiment was basically a repetition of Treisman's but with independent manipulation of prosodic and semantic factors. Pairs of passages of about 50 words each were selected from short stories by H. E. Bates. From each pair of passages, four recordings were made by the same female speaker. Two were of the original passages, and two were made by reading the first part of one passage followed smoothly by the second part of the other. The switch point between the passages was later than halfway through each passage and was always prior to a word beginning with a stop consonant (to facilitate subsequent splicing) but was otherwise placed at random (although never at a major clause or sentence boundary). From these four original recordings, four different dichotic conditions were made: a normal condition in which the two original passages were paired together (aligned to give simultaneity of the stop closures); a semantic change condition, made by pairing together the other two original recordings; an intonation change condition, made by switching the latter pair of passages after the stop closure; and a condition in which both semantics and intonation changed, made by switching the two original passages after the stop closure. The first and last of these conditions repeat Treisman's (1960) conditions, the other two vary independently semantic and intonational continuity on the attended ear. All four conditions were made up with the same number of splicings and rerecordings to prevent the preparation of the tapes selectively introducing artifacts into the experiment.

Each of 14 subjects was instructed to shadow the passage on one ear with as little lag as possible between hearing the speech and saying it. They were told to try not to chunk the speech into phrases before speaking it, and were given five practice trials on similar dichotic passages that did not have any switches. They then took eight different dichotic passages with each of the four experimental conditions appearing twice in a counterbalanced order.

Two types of errors are distinguished in the results. A particular trial is classed as having an omission error if the subject misses at least two words over the break point, and it is classed as an intrusion error if the subject shadowed any words from the unattended ear over the break point. Table 1 shows the distribution of errors over the four different types of break.

TABLE 1: Percentage of trials on which omission or intrusion errors occurred for different types of discontinuity and for subjects who either shadowed continuously or "chunked" their responses.

		Type of discontinuity on shadowed ear			
		No break	Intonation	Semantic	Both
Chunked (5 Ss)	Intrusions	0	20.0	0	0
	Omissions	10.0	40.0	90.0	50.0
Continuous (9 Ss)	Intrusions	0	77.8	5.6	55.6
	Omissions	0	11.1	55.6	22.2
TOTAL (14 Ss)	Intrusions	0	57.1	3.6	35.7
	Omissions	3.6	21.4	67.8	32.1

In the results for all the subjects combined, there is a significant change in the number of intrusion and omission errors across the three experimental conditions ($p<.01$ and $<.05$, respectively). This variation is due to there being significantly more intrusion errors in both the conditions with intonation changed than when only the semantics changed ($p<.05$), and through there being significantly more omission errors in the condition with the semantics alone changed than in the other two experimental conditions ($p<.01$). A closer look at the subjects responsible for the various types of errors showed, in addition, that very few intrusion errors were made by the five subjects, who despite their instructions, chunked their shadowing responses. Omission errors, by contrast, were distributed more evenly between the subjects, but tended to be made more by the subjects who chunked their responses. In other words, for subjects who shadow continuously, an abrupt change of the intonation contour between the ears causes intrusion errors to occur, while an abrupt change in the semantic content of the message, but without any switch in the intonation contour, gives omission errors. Subjects who chunk their shadowing responses likewise produce a large number of omission errors on semantic-switch trials, but show a different pattern from the subjects who shadow continuously on the intonation-switch trials, giving more omission than intrusion errors.

Clearly, the subjects who chunk their responses are much more sensitive to the semantic constraints on the message they are shadowing, in that they show omission errors when there is a semantic discontinuity on the attended ear. Nevertheless they do not show any tendency for intrusions to occur from the other ear when only the semantic information switches ears. Although the subjects who shadow continuously are disrupted less by the semantic discontinuity than their chunking colleagues, they too show no tendency for intrusions to occur from the semantically appropriate unattended ear. Rather than assuming that intrusion errors are caused by some semantic priming of a contextually likely word, it seems, at least in this experiment, that they are occurring only when there is a continuity of intonation across the two ears. For a brief time after a switch in intonation has occurred, the continuity of the intonation contour overrides the ear of entry as a criterion for selecting the speech that needs to be attended. If subjects are shadowing continuously, this leads to intrusion errors from the ostensibly unattended ear. But if the subject is chunking his response, he has time between hearing the potentially intrusive words and making his response to omit the words from the inappropriate ear. This, however, causes omission errors, particularly when there is no semantic continuity on the attended ear to help him retrieve the words that occurred there immediately after the break.

This finding, that prosody helps a listener to attend to a particular speaker, complements the work reviewed earlier on the predictive use of prosody in anticipating stress. Both types of experiment emphasize the role prosody plays in controlling which parts of the speech stream are attended to, whether the selection is made temporally or spatially.

PROSODY AND THE STRUCTURES OF SPEECH

As well as allowing anticipation of the speech stream, prosodic factors doubtless also play a role in segmenting speech into higher-order structures. One rather simple way in which prosody delimits complex structures in speech is through pauses. Goldman-Eisler (1968) claims that even at its most fluent, over two-thirds of speech consists of utterances of six words or less. Pauses virtually never appear within a word, and their length is determined by the type of syntactic juncture that they occur in. They tend to be longer at the end of sentences than at the beginning of subordinate clauses, and before a subordinate clause their length is determined by the type of clause (Goldman-Eisler, 1972). Pauses have been used as experimental variables in studying memory for lists of items that do not naturally fall into larger structures, and we might get some insight into their use in natural language, by looking for a moment at their effect in unnatural situations.

A brief period of silence after an item in the middle of a list of digits provides an opportunity for rehearsal of the preceding group of digits (Kahneman, Onuska, and Wolman, 1968; Ryan, 1969; Kahneman and Wright, 1971), as well as allowing the additional information from an auditory memory to be utilized (Crowder and Morton, 1969). Merely indicating by instructions or interposed tones in a regular list how that list should be grouped gives substantially poorer overall performance and a reduced scalloping of the serial recall curve (Ryan, 1969). This is perhaps because less capacity can be allocated to reheasal when there are competing perceptual demands. As well as delimiting groups of items for rehearsal, pauses also serve to delimit larger structures for coding in long-term memory. The cumulative learning that occurs when supraspan strings of items are repeated occasionally throughout a short-term memory experiment (Hebb, 1961) is reduced when these strings are grouped differently by pauses (Bower and Winzenz, 1969), but this reduction only seems to occur when the items that compose the string are amenable to some higher-level recoding (Laughery and Spector, 1972).

Coming closer to real speech, reading a list of nonsense syllables in a natural intonation can increase the number of syllables remembered immediately after, provided the syllables contain additional bound morpheme clues to the intended syntactic structure (O'Connell, Turner, and Onuska, 1968). Neither intonation, nor the presence of bound morphemes alone is sufficient to increase the number of syllables recalled. Incidentally, both these additional cues also seem to be necessary to procure a right-ear advantage (Zurif and Sait, 1970; Zurif and Mendelsohn, 1972; Zurif, 1974).

It is tempting to extrapolate from these experiments to a view of clause or sentence perception that delays syntactic processing of clausal unit until a clause boundary has been reached, so that the end of the clause is, like the pause in a string of digits, a time of great mental effort (cf. Kahneman et al., 1968; Wright and Kahneman, 1971), but there is some evidence against this. As we remarked earlier, Aaronson (1968) found that reaction time to a target digit in a string that the subject had to remember increased toward the end of the list, reflecting the increased memory load. By contrast, there is very little evidence that a similar increase in reaction time occurs for a phoneme target in a sentence. Foss and Lynch (1969) did find such an increase, but all relevant subsequent studies have found a decrease in reaction time throughout the sentence (Foss, 1969; Hakes and Foss, 1970; Shields et al., 1974). Similarly, reaction time to a click played at the end of a clause decreases with increasing length of the clause, while reaction time to a click presented between clauses shows no consistent increase with length of the first constituent (Abrams and Bever, 1969). Nor is reaction time to a click consistently faster at the beginning of the second clause than in the clause break (Abrams and Bever, 1969; but see Bever and Hurtig, 1975). It is clear then that the constituent words of a clause are not being held in memory like so many digits. Something much more dynamic is happening. But what? Perhaps the best type of model to use here would be one based on Winograd's (1972) integrated language understanding system. In this program both syntactic and semantic information is used, as each successive word in a sentence is encountered, to construct procedures that can subsequently be used to take action on the sentence. The end of a clause does not in this scheme imply a particularly energetic syntactic activity on the part of the computer, provided the syntactic organization that it had presumed during perception of the preceding word string is comparable with the string's ending. Incompatibility requires backtracking, but the program's use of semantic constraints for eliminating alternative syntactic structures as it goes along reduces the likelihood of this being necessary. If these sorts of syntactic and semantic constraints are also being used dynamically in natural perception, then we might suppose that prosody is used in a similarly dynamic and interactive way to guide the search toward an appropriate syntactic organization of the sentence.

Some of the results of an experiment run recently by John Capitman in collaboration with myself and Susan Brady bear a little on these issues. Our experiment was designed primarily to be a replication of Wingfield and Klein's (1971) interesting finding that inappropriate intonation contours impair immediate recall of a sentence.

EXPERIMENT II

The sentences we used in this experiment were similar to those used by Wingfield and Klein (1971). Each sentence that the subjects heard was one member of a group of four sentences derived from pairs of sentences sharing a common string of words. Two of the sentences in the group of four were the originals with appropriate intonation, the other two were generated by cross-splicing the common string of words between these two sentences. This string always started with a stop consonant to facilitate silent splicing, and continued to the end of the sentence to reduce the rhythmic discontinuity that splicing can introduce. The sentences were between 10 and 14 words long and the common string was between 6 and 10 words long. Unlike the Wingfield and Klein material, the major syntactic boundary in each sentence was both preceded and followed by at least two words of shared material, this ensured that the intonationally suggested boundary in the mosaic sentences generated by cross-splicing was also preceded and followed by at least two words of shared material. As in the Wingfield and Klein experiments the sentences were presented monaurally and the ear of presentation was switched within the sentence. This switch was always made during a stop closure to prevent extraneous clicks. The switch point always occurred within the shared material and was never at the intonationally suggested boundary, but was otherwise unconstrained. The recordings were made on the Haskins Laboratories pulse-code-modulation (PCM) facility.

Each of 24 subjects heard one sentence of the four generated from each of the 11 pairs in an order that counterbalanced conditions (normal versus cross-spliced). Subjects were instructed to write down each sentence as soon as they had heard it. In order to make the task more difficult, they were told to listen for the occurrence of stop consonants and to circle these in their answer. Their performance on this task was ignored.

On average, 55.5 percent of the normally intonated (unspliced) sentences were recalled correctly, while only 44.5 percent of the cross-spliced sentences were recalled correctly. In examining the errors that subjects made, four types of error are distinguished:

1. Omission errors: Each word omitted, whether or not this omission changed any other aspect of the sentence, was scored as an omission error.

2. Lexical errors: Each changed word in a string that did not affect the primary grammatical relations (the truth conditions of the sentence) was scored as a lexical error. Changes in tense, mood, and aspect were also counted as lexical errors.

3. Syntactic errors: Each change in any segment of a string that resulted in a change in the grammatical relations or the truth conditions of the sentence was scored as a syntactic error (except for errors of tense, mood, and aspect). No attempt was made to distinguish those changes that seemed to conform to the intonation pattern from those that did not. One other error type was defined that depended on the position of the error but not on its type:

4. IB2 errors: [This score applied only to the cross-spliced sentences.] Each error of any type occurring within one word on either side of the intonationally suggested boundary was scored as an IB2 error.

Summing together all the first three types of errors (omission, lexical, syntactic), Capitman found significantly more errors on the cross-spliced than on the normally intonated sentences. This difference is significant by a Wilcoxon test both across subjects ($p<.01$) and across sentences ($p<.02$). Moreover, this increase in errors is due to an increase in the omission ($p<.02$ by subject; $p<.05$ by sentence) and syntactic errors ($p<.01$ by both subject and sentence), but not to an increase in lexical errors. An exception is seen in three of the sentence pairs that show a decrease in syntactic errors in the crossed-intonation condition, though still an increase in omission errors. All of these sentences have an "If...then..." construction, which may impose greater syntactic or semantic constraints on the subjects' responses, causing them to give up on their response rather than change the truth conditions of the sentence.

Although these general results confirm Wingfield and Klein's (1971) findings on intelligibility with a different set of sentences, they do not get us much nearer to the question of how the inappropriate intonation contour influences recall. Inspection of the IB2 errors does bear on this question. When listening to the crossed-intonation sentences, subjects make more errors within one word on either side of the intonationally suggested boundary when this boundary precedes the major syntactic boundary than when it follows it ($\underline{p}<.025$, by subject; $\underline{p}<.05$ by sentence). This is not owing to subjects' making more errors earlier in the sentence, since the reverse tends to be the case. This asymmetry in the disruptive effects of an inappropriate intonational cue to a clause boundary is what one might expect on the hypothesis that intonational information is being used dynamically to restrict syntactic hypotheses as the subject listens to the sentence. If prosody were being used solely to restrict alternatives after initial perception of the sentence, it is difficult to see why the relative positions of the syntactic and intonational boundaries should matter. On a more dynamic view, it is possible that the inappropriate intonational boundary leads to backtracking throughout the previous incomplete clause; this clause will be much longer, and so the effects of backtracking will be more disruptive when the incomplete clause is the first one of the sentence. Hence, having the intonational boundary before the syntactic boundary will be more disruptive than having it after it. Looking at errors only in the region of the intonation boundary appears to be a more sensitive measure than looking at errors over the whole sentence, since there did not appear to be any significant difference in errors over the sentence as a whole as a function of these relative locations. This perhaps suggests that backtracking affects perception of the part of the sentence that is being actively perceived when it occurs. But we cannot entirely rule out the possibility that this effect is being caused in part by the rhythmic discontinuity that cross-splicing introduces immediately after the breakpoint. Although these conclusions are extremely speculative and the evidence on which they are based is not strong, we feel that the questions raised by this sort of approach are extremely interesting, and we hope to pursue them in subsequent experiments.

In summary, this paper has tried to draw attention to some of the possible dynamic roles that prosody might play in the perception of speech: the rhythmic and melodic aspects of speech may allow the listener to predict when potentially important speech material will arrive (and perhaps allow him to allocate his processing capacity accordingly); they may also allow him to attend selectively to one voice among many more readily. In addition, prosody undoubtedly plays a role in delimiting higher-order structures in speech, and the suggestion is made here that this is done by dynamically modifying the hypotheses the listener entertains while listening to a sentence. While we neither pretend that this exhausts the uses of prosody, nor believe that the study of isolated sentences is the most appropriate way to study a variable that depends so much on intersentence context (see Smith and Goodenough, 1971, for an interesting example of intonation interacting with context in a perceptual task), we do hope that the issues raised here will help to stimulate work on an unduly neglected area of speech perception.

REFERENCES

Aaronson, D. (1966) Temporal course of perception in an immediate recall task. *J. Exp. Psychol. 76*, 129-140.

Abrams, K. and T. G. Bever. (1969) Syntactic structure modifies attention during speech perception and recognition. *Quart. J. Exp. Psychol. 21*, 280-290.

Allen, G. D. (1972) The location of rhythmic stress beats in English: An experimental study, I and II. *Lang. Speech 15*, 72-100, 179-195.

Bever, T. G. and R. R. Hurtig. (1975) Detection of a nonlinguistic stimulus is poorest at the end of a clause. *J. Psycholing. Res. 4*, 1-7.

Blesser, B. (1969) Perception of spectrally rotated speech. Unpublished Ph.D. dissertation, Massachusetts Institute of Technology.

Bower, G. H. and D. J. Winzenz. (1969) Group structure, coding and memory for digit series. *J. Exp. Psychol. 80*, no. 2, part 2.

Bregman, A. S. and J. Campbell. (1971) Primary auditory stream segregation and perception of order in rapid sequences of tones. *J. Exp. Psychol. 89*, 244-249.

Cairns, H. S. and D. J. Foss. (1971) Falsification of the hypothesis that word frequency is a unified variable in senence processing. *J. Verbal Learn. Verbal Behav. 10*, 41-43.

Crowder, R. G. and J. Morton. (1969) Pre-categorical acoustic storage (PAS). *Percept. Psychophys. 5*, 365-373.

Cutler, A. (1975) Rhythmic factors in the determination of perceived stress. Paper presented to the 89th meeting of the Acoustical Society of America, April, Austin, Tex.

Cutler, A. and D. J. Foss. (1973) The importance of lexical item stress for lexical access. Paper presented at the 44th annual meeting of the Midwestern Psychological Association, 10-12 May, Chicago.

Dooling, D. J. (1974) Rhythm and syntax in sentence perception. *J. Verbal Learn. Verbal Behav. 13*, 255-264.

Dorman, M. F., J. Cutting, and L. J. Raphael. (1975) Perception of temporal order in vowel sequences with and without formant transitions. *J. Exp. Psychol.: Human Perception and Performance 2* (now vol. 1), 121-129.

Foss, D. J. (1969) Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *J. Verbal Learn. Verbal Behav. 8*, 457-462.

Foss, D. J. and R. H. Lynch. (1969) Decision processes during sentence comprehension: Effects of surface structure on decision times. *Percept. Psychophys. 5*, 145-148.

Foss, D. J. and D. A. Swinney. (1973) On the psychological reality of the phoneme: Perception, identification and consciousness. *J. Verbal Learn. Verbal Behav. 12*, 246-257.

Fry, D. B. (1970) Prosodic phenomena. In *Manual of Phonetics*, ed. by B. Malmberg. (Amsterdam: North-Holland).

Fujimura, O. (1961) Some synthesis experiments on stop consonants in the initial position. *Quarterly Progress Report* (Research Laboratory of Electronics, MIT) *61*, 153-162.

Goldman-Eisler, F. (1968) *Psycholinguistics: Experiments in Spontaneous Speech*. (London: Academic).

Goldman-Eisler, F. (1972) Pauses, clauses, sentences. *Lang. Speech 15*, 114-121.

Haggard, M. P., S. Ambler, and M. Callow. (1970) Pitch as a voicing cue. *J. Acoust. Soc. Amer. 47*, 613-617.

Hakes, D. T. (1972) Effects of reducing complement constrictions on sentence comprehension. *J. Verbal Learn. Verbal Behav. 11*, 278-286.

Hakes, D. T. and H. S. Cairns. (1970) Sentence comprehension and relative pronouns. *Percept. Psychophys. 8*, 5-8.

Hakes, D. T. and D. J. Foss. (1970) Decision processes during sentence comprehension: Effects of surface structure reconsidered. *Percept. Psychophys. 8*, 413-416.

Hebb, D. O. (1961) Distinctive features of learning in the higher animal. In *Brain Mechanisms and Learning*, ed. by J. F. Delafresneye. (Oxford: Blackwell).

Huggins, A. W. F. (1972) On the perception of temporal phenomena in speech. *J. Acoust. Soc. Amer. 51*, 1279-1290.

Kahneman, D., L. Onuska, and R. E. Wolman. (1968) Effects of grouping on the pupillary response in a short-term memory task. *Quart. J. Exp. Psychol. 20*, 309-311.

Kahneman, D. and P. Wright. (1971) Changes of pupil size and rehearsal strategies in a short-term memory task. *Quart. J. Exp. Psychol. 23*, 187-196.

Laughery, K. R. and A. Spector. (1972) The roles of recoding and rhythm in memory organization. *J. Exp. Psychol. 94*, 41-48.

Lieberman, P. (1963) Some effects of semantic and grammatical context on the production and perception of speech. *Lang. Speech 6*, 172-187.

Martin, J. G. (1972) Rhythmic hierarchical versus serial structure in speech and other behavior. *Psychol. Rev. 79*, 487-509.

O'Connell, D. C., E. A. Turner, and L. A. Onuska. (1968) Intonation, grammatical structure, and contextual association in immediate recall. *J. Verbal Learn. Verbal Behav. 7*, 110-116.

Rubin, P., M. Turvey, and Van Gelder. (in press) Initial phonemes are detected faster in spoken words than in nonwords. *Haskins Laboratories Status Report on Speech Research SR-44*.

Ryan, J. (1969) Grouping a short-term memory: Different means and patterns of grouping. *Quart. J. Exp. Psychol. 21*, 137-147.

Shields, J. L., A. McHugh, and J. G. Martin. (1974) Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *J. Exp. Psychol. 102*, 250-255.

Smith, F. and C. Goodenough. (1971) Effects of context, intonation, and voice on the reaction time to sentences. *Lang. Speech 14*, 241-250.

Stowe, A. N. and D. B. Hampton. (1961) Speech synthesis with pre-recorded syllables and words. *J. Acoust. Soc. Amer. 33*, 810-811.

Svennson, S.-G. (1974) Prosody and grammar in speech perception. *MILUS* no. 2. (Institute of Linguistics, University of Stockholm).

Treisman, A. M. (1960) Contextual cues in selective listening. *Quart. J. Exp. Psychol. 12*, 242-248..

Treisman, A. M. and J. Tuxworth. (1974) Immediate and delayed recall of sentences after perceptual processing at different levels. *J. Verbal Learn. Verbal Behav. 13*, 38-44.

Wingfield, A. (1975) Acoustic redundancy and the perception of time-compressed speech. *J. Speech Hearing Res. 18*, 139-147.

Wingfield, A. and J. F. Klein. (1971) Syntactic structure and acoustic pattern in speech perception. *Percept. Psychophys. 9*, 23-25.

Winograd, T. (1972) *Understanding Natural Language*. (New York: Academic Press).

Wright, P. and D. Kahneman. (1971) Evidence for alternative strategies of sentence retention. *Quart. J. Exp. Psychol. 23*, 197-213.

Zurif, E. B. (1974) Auditory lateralization: Prosodic and syntactic factors. Brain Lang. 1, 391-404.

Zurif, E. B. and M. Mendelsohn. (1972) Hemispheric specialization for the perception of speech sounds: The influence of intonation and structure. Percept. Psychophys. 11, 329-332.

Zurif, E. B. and P. E. Sait. (1970) The role of syntax in dichotic listening. Neuropsychologia 8, 239-244.

Speech and the Problem of Perceptual Constancy[*]

Donald Shankweiler,[+] Winifred Strange,[++] and Robert Verbrugge[+++]

ABSTRACT

Speech signals are intrinsically variable for many reasons. In this paper we consider the implications of variability for a theory of vowel perception. Current theories of the vowel emphasize the relational nature of the acoustic cues since no absolute values of formant frequencies could unambiguously distinguish vowels produced by different talkers and in different phonetic contexts. It has been assumed that the perceptual process of vowel identification includes a normalization stage whereby the listener calibrates his perceptual apparatus for each talker, according to some reference derived from preceding utterances by that talker. We have been unable to obtain evidence for such a perceptual mechanism. Theories of vowel perception have failed to give due weight to the richness of the natural speech signal. We attempt to show why the invariant acoustic information that specifies a vowel cannot be found in a temporal cross section, but can only be specified over time.

---

[*] Chapter prepared for Perceiving, Acting, and Comprehending: Toward an Ecological Psychology, ed. by R. Shaw and J. Bransford. (Hillsdale, N. J.: Lawrence Erlbaum Assoc., in press).

[+] Also University of Connecticut, Storrs.

[++] University of Minnesota, Minneapolis.

[+++] Human Performance Center, University of Michigan, Ann Arbor.

Acknowledgment: The authors' research reported in this paper was begun during the academic year 1972-73, while D. Shankweiler was a guest investigator at the Center for Research in Human Learning, University of Minnesota, Minneapolis. The early portion of the work was supported in part by a grant from the National Institute of Child Health and Human Development (NICHD) to the Center, and in part by grants awarded to Shankweiler and to J. J. Jenkins by the National Institute of Mental Health. The later portion of the research and the preparation of this paper was supported by grants from NICHD to the Center and to Haskins Laboratories.

We are grateful to T. Edman for his substantial assistance with all phases of the experimental work, and to J. J. Jenkins for advice and encouragement from the project's inception. Other colleagues who commented on earlier drafts of this paper are: G. M. Kuhn, T. Nearey, and M. Studdert-Kennedy. It is a pleasure to acknowledge their help.

[HASKINS LABORATORIES: Status Report on Speech Research SR-42/43 (1975)]

Speech perception is not ordinarily included in the body of phenomena and theory that convention defines as the psychology of perception. Yet the problem of how the perceptual categories of speech are specified in the acoustic signal is a primary example of the problem of perceptual constancy. In spite of its neglect by psychology, speech and its signal have been intensively studied by members of other disciplines. We think that some of the results and puzzles generated by this research are relevant to the concerns of our colleagues whose primary interests are in other facets of human cognition.

Speech perception at any level involves classification. The classificatory step is assumed whenever we move beyond a purely physical (acoustic) description of speech to a psychophysical description in terms of perceptual units. Unlike certain problems in traditional psychophysics in which the choice of units may be arbitrary, there is wide consensus about what the units of perception are in the case of speech. This consensus is the product of centuries of linguistic investigation, during which many attempts have been made to isolate the various levels and units that constitute our perception of speech. Viewed in terms of structure, speech is a hierarchical system that manifests what Hockett (1958) called a "duality of patterning": it employs both meaningful and meaningless units. Morphemes (or, roughly speaking, words) are the smallest of the meaningful units. In all languages morphemes have an internal structure composed of smaller meaningless segments, the phonemes (Bloomfield, 1933). Since the communication of meanings ultimately rests on a foundation of phonemic structure, a basic part of the task of understanding how speech is perceived is to discover the conditions for the perception of phonemic categories.[1] For present purposes, we shall ignore meaning and concern ourselves only with the phonemic message--that is, with the perception of syllables and their phoneme segments, familiar to us as the consonants and vowels.

In speech, as in handwriting, no two "signatures" are alike. In generating the "same" phoneme, different speakers do not produce sounds that are acoustically the same. Indeed, the same signal is never exactly repeated by the same speaker. In perceiving speech, as in identifying objects, we ordinarily regard only those distinctions that are critical, ignoring those that are merely incidental. While no one would deny that speech signals are intrinsically variable for many reasons, the implications of this variability for perception have not been widely appreciated. In brief, they constitute a major problem in *perceptual constancy*.

---

[1] The phoneme is the minimal unit by which perceivers differentiate utterances. For example, the word *bad* has three phonemic segments, /b/, /æ/, and /d/, that differentiate it from such words as *dad*, *bed*, and *bat*. In different utterances, a phoneme may be realized acoustically in different ways; linguists call these variants "phones." For example, the final /t/ in *bat* might be either released (acoustically, a pause followed by a burst) or unreleased (no burst). The class of phones is potentially infinite, and it is arguable whether phones (however defined) are natural perceptual units. Our emphasis in this paper is on how the identity of a phoneme is perceived despite variations in its acoustic form.

THE PERCEIVING MACHINE AND THE ONE-TO-ONE PROBLEM

Although our concern is with how the human perceptual system works, it may help us to bring this problem into focus if we consider how a machine might proceed in recovering a string of phonemic segments. Consider, for example, the problems to be solved in designing a voice-operated typewriter. The goal of such an automatic speech recognition device is to type out the appropriate string of phonemic symbols (or perhaps standard orthographic symbols) in response to any speech input. In the simplest case, the only information available to the device will be the acoustic waveform itself. A human listener, of course, can usually take advantage of other sources of information, including both the linguistic and the situational context of the utterance.[2] While acknowledging the importance of context, we should not overlook the fact that listeners can identify arbitrarily chosen words and nonsense syllables with high accuracy when listening conditions are favorable. In other words, we are not posing an unrealistic problem for our hypothetical device; human listeners can do remarkably well when little contextual information is available.

Many attempts have been made in recent decades to design a voice-operated typewriter, but the problem has so far proved elusive. Despite a degree of success with severely restricted vocabularies when words are spoken by a trained talker, a generally useful speech recognizer continues to be unattainable. As Hyde (1972:399) notes, "there are still no devices which can perform even moderately well on normal (conversational) speech in normal (noisy) environments by a normal range of talkers."

It is worth considering for a moment how the operation of a speech recognition system has typically been conceived. As in many other automatic pattern recognition devices, the procedure involves two stages. In the first stage, the basic units or segments are located. For example, in automatic reading of print, this would correspond to isolating individual letters. In the second stage, each segment is identified as an instance of one of a fixed set of objects. In the case of print reading, this would correspond to identifying a segment as a particular letter of the alphabet. Thus a successful voice-operated typewriter would have to be able to perform two operations on the acoustic waveform of any speech signal. First, it would have to divide the waveform into acoustic segments that have a one-to-one correspondence through time with the sequence of phonemes in the utterance. Second, it would have to detect the presence or absence of acoustic features that are critical for identifying particular phonemes. This second stage is often conceived of as requiring a set of filters, each filter being tuned to a critical acoustic property (defined along dimensions such as frequency, intensity, and duration).

This strategy implies certain widespread assumptions that only in recent years have been successfully challenged. For example, we find that in the standard accounts of speech acquisition, it is tacitly assumed that speech consists of a collection of elementary sounds "transparent" to the infant, such that he automatically recognizes a parent's utterance of /d/ as "the same sound" as his own utterance of /d/ (Allport, 1924; Watson, 1924). Similarly, taxonomic linguists working in the tradition of Bloomfield (1933) supposed that all languages sampled from a common inventory of sounds (phones). Working from phonetic transcription of a large number of utterances as a base, these linguists developed highly successful procedures for determining which sound contrasts played a role in any particular language. It was believed that the great practical success of transcription as a tool for language description rested on a narrow physical base, and that, in principle, an acoustic definition could be given for each phone. In this view, speech was conceived as a kind of sound alphabet, in which each phone is conveyed by a discrete package of sound with a characteristic spectral composition. The pervasiveness of this assumption has been noted by Denes (1963:892):

---

[2] The importance of context to human listeners has been elegantly demonstrated by Miller, Heise, and Lichten (1951), who found a remarkably predictable relationship between the amount of acoustic distortion that yields a given level of intelligibility and the informational redundancy of the message. For a given signal-to-noise ratio, intelligibility was greater for words heard in sentence context than for words heard in isolation.

> The basic premise of [most speech-recognition] work has always been that a one-to-one relationship existed between the acoustic event and the phoneme. Although it was recognized that the sound waves associated with the same phoneme would change according to circumstances, there was a deep-seated belief that if only the right way of examining the acoustic signal was found, then the much sought-after one-to-one relationship would come to light.

In fact, the perceptual skills that underlie phonetic transcription have never been explained well enough that an algorithm could be written to permit a machine to do the job. From our present perspective it is clear why no one has been able to develop a voice-operated typewriter based on the startegy outlined above. First, there are no clearly bounded segments in the acoustic waveform of roughly phonemic size; that is, there are no acoustic units available for setting up a correspondence with phonemes. Second, even if boundaries are arbitrarily imposed on the continuous signal, the segments corresponding to a particular phoneme often vary considerably in their acoustic composition. Moreover, any one of those acoustic segments, transferred to a different phonemic environment, might be heard as a different phoneme altogether. Not only do the physical attributes specifying a particular phoneme vary markedly, but the same physical attribute can specify different phonemes depending on the context.

THE CONTINUOUS SIGNAL DOES NOT REVEAL THE SEGMENTATION OF THE PHONEMIC MESSAGE

The conclusions stated above are the results of three lines of investigation begun in the mid 1940s and continuing to the present. We turn now to review briefly the nature of the evidence.

Of special importance from our standpoint are a series of tape-cutting and tape-splicing experiments that had the effect of shaking the general confidence that phonemes are conveyed by isolable bits of sound. These experiments failed to find any way to divide the signal on the time axis to yield segments of phoneme size. For example, given a consonant-vowel syllable such as go, there is no way to cut the piece of magnetic tape so as to produce the consonant /g/ alone. Some vowel quality always remains. Moreover, if a consonant-vowel syllable is cut at some point, the consonantal portion may not be heard as the same phoneme when spliced to a recording of a different vowel. Schatz (1954), for example, found that the consonantal portion of an utterance of /pi/ was heard as /k/ when it was joined to the vowel /ɑ/. Harris (1953) and Peterson, Wang, and Sivertsen (1958) independently concluded that assembled speech made by splicing together prerecorded segments is not generally intelligible when the units are smaller than roughly a half-syllable.[3] Some investigators (e.g., Cole and Scott, 1974a, 1974b) continue to argue that speech perception may be based in large part on the detection of acoustic invariants for phonemes. They have claimed that a one-to-one correspondence can often be found, that spliced segments can preserve their identity when transferred to new phonemic contexts. However, much of this apparent "invariance" disappears when one is careful to cut the initial segment sufficiently short that no trace of the subsequent vowel remains (cf. Kuhl, 1974).

A second important development is the study of spectrographic displays of speech. This work was made possible by the invention of the sound spectrograph (Koenig, Dunn, and Lacey, 1946) during World War II and by the general availability of such devices for research during the postwar years. The spectrograph displays, in graphical form, the time variations of the spectrum of the speech wave. This representation of the sound patterns of speech is valuable for the information it gives about articulation. The energy in speech sounds is concentrated in a small number of frequency regions that appear on a spectrogram as horizontal bands (called "formants"). The location of the formants on the frequency scale reflects the primary resonances of a talker's vocal tract (Fant, 1960). Since the shape of the vocal cavity changes at the joining of successive consonants and vowels, the formant frequencies may be seen to modulate up and down as one scans a spectrogram along the time axis. However, efforts to locate discrete information-bearing units along the time axis have met with repeated failure. The phonemic and syllabic segments, which are so clear perceptually, have no obvious correlates in a spectrogram, as evidenced by the fact that spectrograms are very difficult to read, even after much experience (Fant, 1962; but see also Kuhn and McGuire, 1975).[4]

Figure 1 shows spectrograms of two syllables, bib and bub. Note that the formant frequencies are nonoverlapping for the entire duration of the syllables, not just in the middle portion. Although the syllables differ phonemically (i.e., perceptually) in only one segment (the medial vowel), acoustically they differ throughout.

Failure to find obvious acoustic cues in spectrograms led to a third line of investigation: a variety of experiments with synthetic speech, produced by devices that place acoustic parameters under the experimenter's direct control. Early work by researchers at Haskins Laboratories made use of hand-painted

---

[3] It does not follow from this result that the minimal perceptual unit is larger than the phoneme. The inference we would draw is that decisions about phoneme identity are made with regard to information distributed over the whole syllable (and sometimes, perhaps, over a number of syllables).

[4] In remarking on the difficulty of reading spectrograms our point is not that the spectrogram does not represent the relevant phonemic information, but rather that the ear has readier access to the brain's phonemic decoder than the eye.

FIGURE 1

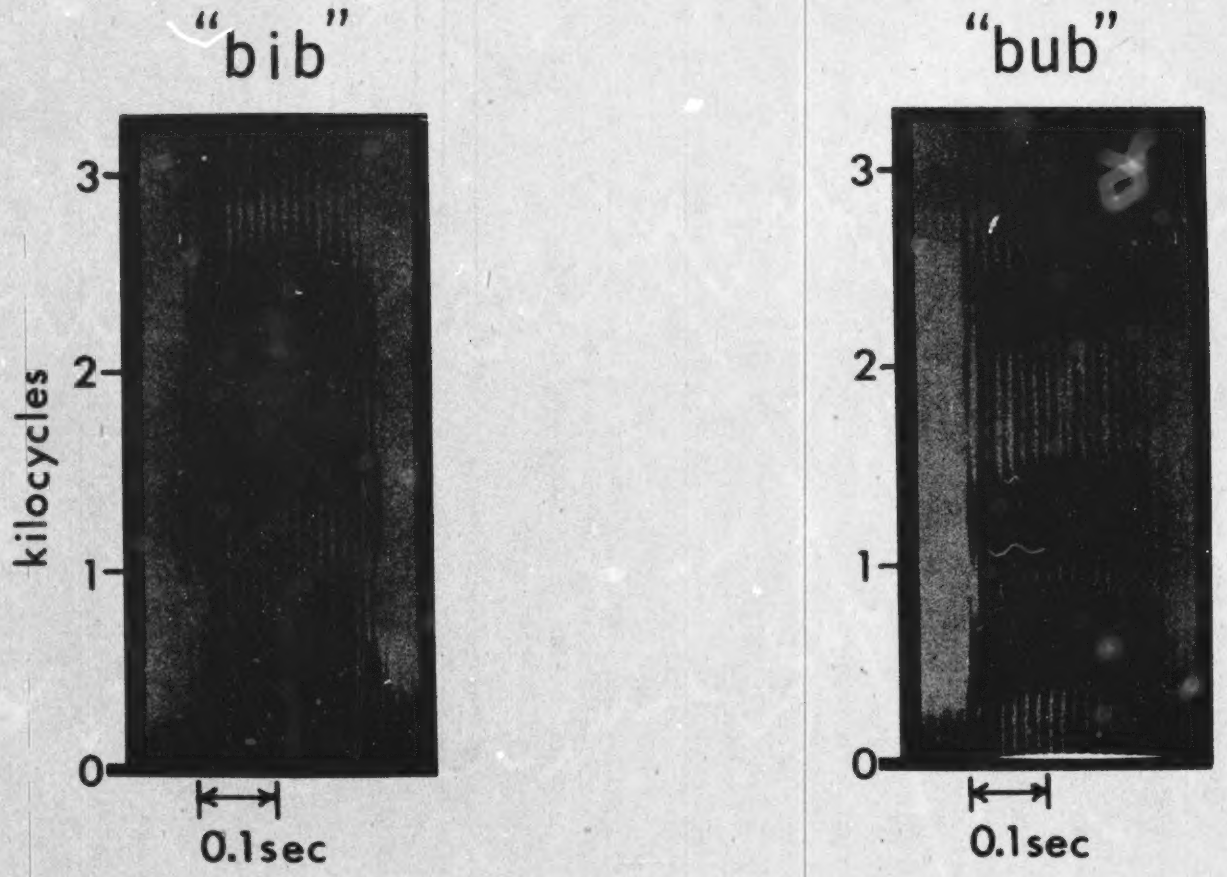

Figure 1: Spectrograms of tokens of two syllables that differ in the vowel. Note that the formant pattern of "bib" does not coincide with that of "bub" in any portion.

patterns resembling spectrograms that were converted into sound by a photoelectric device, the Pattern Playback (Cooper, Liberman, and Borst, 1951). The Pattern Playback and subsequent computer-controlled electronic synthesizers have made it possible to do analytic studies in which one parameter is varied at a time, to determine which parameters were critical for particular phonemes. Only through systematic psychophysical experimentation of this sort has it been possible to locate the linguistically relevant information in the speech spectrum (Cooper, Delattre, Liberman, Borst, and Gerstman, 1952; Liberman, 1957; Liberman, Harris, Hoffman, and Griffith, 1957; Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). A major conclusion of this research is that, in general, there is no simple one-to-one correspondence between perceptual units and the acoustic structure of the signal. To be successful, synthetic speech must encode the information for phonemes into acoustic patterns at least a half-syllable or full-syllable in length.

These findings make it possible to understand why the design of a voice-operated typewriter proved so difficult. Phonemes are not merely joined acoustically; they overlap so that two or more are represented simultaneously on the same stretch of sound. Conversely, segmentation is impossible because information for one phoneme is usually spread over wide stretches of the signal. Even if segmentation were attempted, the cues isolated would be radically different in another phonemic environment. As we saw in Figure 1, all four segments corresponding to /b/ would differ markedly in slope and formant frequency range.

In sum, the radically context-dependent structure of speech dooms to failure the kind of pattern recognition procedure outlined above. A procedure that combines a prior stage of segmentation with an analysis of segments by a set of tuned-filter "detectors," operating independently and in parallel, will be insufficient as a job description of an automatic recognition machine (and insufficient as a model of human speech perception as well). One contemporary approach to the recognition problem (Mermelstein, 1974) is explicit on this point, acknowledging that a speech recognizer would have to extract information about component phonemes over longer stretches of the signal than a syllable and would have to incorporate rules about how that information is distributed. The system described by Mermelstein does not assume that the segmentation and labeling problems are independent.

SYLLABLE NUCLEI AS TARGETS

A commonly suggested strategy for speech recognition (Fant, 1970) is to classify first those phonemes that are most "transparent" in the signal, and then to use that information as a basis for determining what the more contextually variable phonemes are. For example, the research described earlier found that the acoustic form of many consonants is heavily dependent on the coarticulated vowel. If the vowel were easily detected in speech signals, then it could be identified first and used as a basis for disambiguating the neighboring consonant.

There are several reasons for supposing that vowels could be extracted readily by a routine based on a filter bank, though it would not be possible, in general, for consonants. In productions of sustained vowels, the positions of the formants on the frequency axis serve roughly to distinguish the vowels of a particular talker. For a high front vowel such as /i/, the first and second formants are spaced far apart (on the frequency axis) while the second and third formants are spaced close together. For a high back vowel /u/, the pattern is just reversed. These patterns can be synthesized with a combination of steady-state resonances that simulate the formants found in spectrograms of natural vowels. Synthetic stimuli are readily labeled as vowels by listeners, and it is possible to generate the full complement of English vowels with two or three formants (Delattre, Liberman, Cooper, and Gerstman, 1952).

This acoustic characterization of vowels as steady-state entities is reinforced by articulatory considerations. The flow of speech is marked by a rhythmic pattern of syllables; each syllable contains a vowel "nucleus" that is usually coarticulated with one or more consonants. It is usual to think of consonants as the dynamic component of speech, since they are generally produced by movement of the articulators, and to regard vowels as the static component, since they may be produced with a stationary vocal-tract configuration and sustained indefinitely.

This contrast is emphasized by the concept of an idealized vowel as a prolonged, static entity defined (acoustically) by the frequencies of the first two or three formants, i.e., by the primary resonances of the stationary vocal tract. At least for individual talkers, then, there should be a distinctive set of frequency values associated with each vowel, in contrast to the variable values associated with the talker's consonants. In that case, vowels should be retrievable by a simple two-stage recognition procedure: it would be a straightforward matter to detect the presence of steady states electronically (thereby isolating vowel segments for analysis), and a filter bank could then determine which set of critical frequencies the vowel sound best fits.

Unfortunately for this approach, the apparent simplicity of vowels is largely an illusion. Vowels in natural continuous speech, unlike the artificially prolonged vowels of the phonetics laboratory, are not generally specified by steady states at all. Let us see why this is so.

First, as a result of coarticulation, vowels are encoded into the structure of a full syllable. The imprint of the vowel is not localized but is smeared throughout the entire temporal course of the syllable. Thus, information about a vowel is available in the transitions as well as in the steady-state portion (if, indeed, a steady state is even attained). This was clear in the earlier example of the syllables bib and bub; in both cases, the vowel affected the spectral pattern of the entire syllable. Moreover, the acoustic properties of the vowel nucleus may be affected by coarticulated consonants. Measurements by House and Fairbanks (1953) and Stevens and House (1963), for example, indicated consistent changes in the duration, fundamental frequency, and formant frequencies and intensities of vowels, depending on consonantal context. It should also be noted that consonants can affect the structure of neighboring vowels if (by the phonological rules of the dialect) a distinction between two consonants is actually manifested by a difference in the neighboring vowels. For example, the /d/ in rider is distinguished from the /t/ in writer by the increased duration of the vowel that precedes it. Similarly, in some dialects of English, a nasal phoneme, such as the /n/ in pants, is realized by a nasalization of the preceding vowel; thus the spectral structure of the vowel /æ/ will vary markedly depending on whether pats or pants is spoken. Because the coarticulation effects between consonants and vowels do not operate in one direction alone, but are two-way effects, there is no obvious acoustic invariant that characterizes a vowel in all consonantal contexts.

A second major source of variance in the acoustic structure of vowels is the tempo of articulation. During rapid rates of speech, steady-state configurations may never be attained at all. Acoustic analysis of rapid speech supports the hypothesis of articulatory "undershoot," since syllable nuclei often do not reach the steady-state formant frequency values characteristic of vowels in slowly articulated syllables (Lindblom, 1963; Stevens and House, 1963). Lindblom and Studdert-Kennedy (1967) found that listeners showed a shift in the acoustic criteria that they adopted for vowels (i.e., there was a shift in the phoneme boundary between them) as a function of perceived rate of utterance. Apparently, human listeners compensated for this simulated articulatory undershoot by perceptual overshoot. These data show that formant transitions, which are generally understood to carry consonantal information, may also aid in specifying the vowel. Thus, in ordinary speech, vowels, like consonants, are dynamic entities that are scaled by the pace of speaking.

A third source of acoustic variation in vowels is associated with the individual characteristics of the talker. We perceive this variation directly when we identify persons on the basis of voice quality. On the other hand, such individual variation is irrelevant and becomes "noise" when our intent is to recover the linguistic message. Inasmuch as formant frequencies reflect vocal-tract dimensions, it is obvious that the absolute positions of the formants will not be the same for a child as they are for an adult. The extent of the problem is suggested by Joos (1948:64).

> The acoustic discrepancies which an adult has to adjust for when listening to a child speaker are nothing short of enormous--they commonly are as much as seven semitones or a frequency ratio of 3 to 2, about the distance from /ɛ/ to /ʊ/.

Somehow, in spite of this, we manage to understand small children's speech reasonably well and they ours. This is especially remarkable given that the difference between children and adults cannot be described by a simple scale factor. The vocal tract not only increases in size but changes in shape, and the consequent changes in the acoustic output are correspondingly complicated. Indeed, Fant (1966) has argued that the assumption of an invariant relation between formants one and two for a given vowel is just as untenable as the assumption that the absolute formant frequencies of the vowel are invariant for all talkers. Thus, the relation between utterances of a syllable by an adult and a small child is not the multiplicative relation that obtains between versions of a melody played in different keys.

Having discussed variation based on physical differences in the sound-production apparatus, we should also mention differences that are social in origin, reflecting local variations of dialect within the larger language community. These variations are associated, of course, with geographical region, ethnic group, and socioeconomic class. Additional sources of talker-related variation are idiosyncratic speech mannerisms, emotional state, and fatigue. These sources, in addition to those we have discussed above, pose enormous difficulties to the design of an automatic recognition device.[5]

No one, to our knowledge, has seriously considered how an automatic speech recognition routine would adjust its criteria to compensate for the variations associated with coarticulation, tempo, and talker. When we consider the magnitude and variety of variations that we take in our stride as perceivers, we begin to realize something of the complexity of the relations between the signal and the phonemic message. The difficulties encountered by the task of machine recognition command a new respect for the subtlety and versatility of the human perceptual apparatus and lead us to a new appreciation of the abstract nature of speech perception.

WHAT SPECIFIES A VOWEL?

The idea that vowels can be defined as fixed sets of steady-state values is an oversimplification that bears little relation to the structure of natural speech. We have found it necessary to reopen the question of what specifies a vowel, and we wish to introduce some recent findings as a case study in the problem of perceptual constancy.

Vowels, as we noted earlier, are traditionally defined by formants. Vowel quality is associated with concentrations of acoustic energy in a few relatively narrow portions of the frequency spectrum; energy in the regions between these bands is generally weak and has little perceptible effect on vowel quality. In distinguishing among vowels, the lowest two formants are traditionally thought to be the most significant; the contribution to perception of the third and higher formants is problematical. For this reason, vowels are customarily represented as points located in a two-dimensional space defined by the first and second formants. As a result of variations among talkers, the points in this acoustic vowel space are actually regions. A critical question for perceptual theory is: How much or how little do these regions overlap?

A thorough assessment of this question was made by Peterson and his colleagues (Peterson, 1951; Peterson and Barney, 1952), who obtained spectrographic measurements of tokens of 10 American English vowels produced by 76 talkers (including men, women, and children). Figure 2, which is redrawn from Peterson and Barney (1952), shows the vowel space defined by measurements of formants one and two (F_1 and F_2). We note that there is considerable overlap in some regions. In running speech we might expect a comparable analysis to show still more overlap. The findings showed not only lack of invariance in the position of the formants in children and adults, but also considerable average differences between men and women and considerable variation among talkers of the same age group and sex.

In his pioneering monograph on acoustic phonetics, Joos (1948) had discussed the dilemma that such variation poses for theories of speech perception. If

---

[5] Reflect, too, on the variety of transformations of the signal that might be produced by the commonplace feats of talking with food in the mouth, with a cigar between the lips, or with teeth firmly clamped on a pencil (cf. Nooteboom and Slis, 1970).

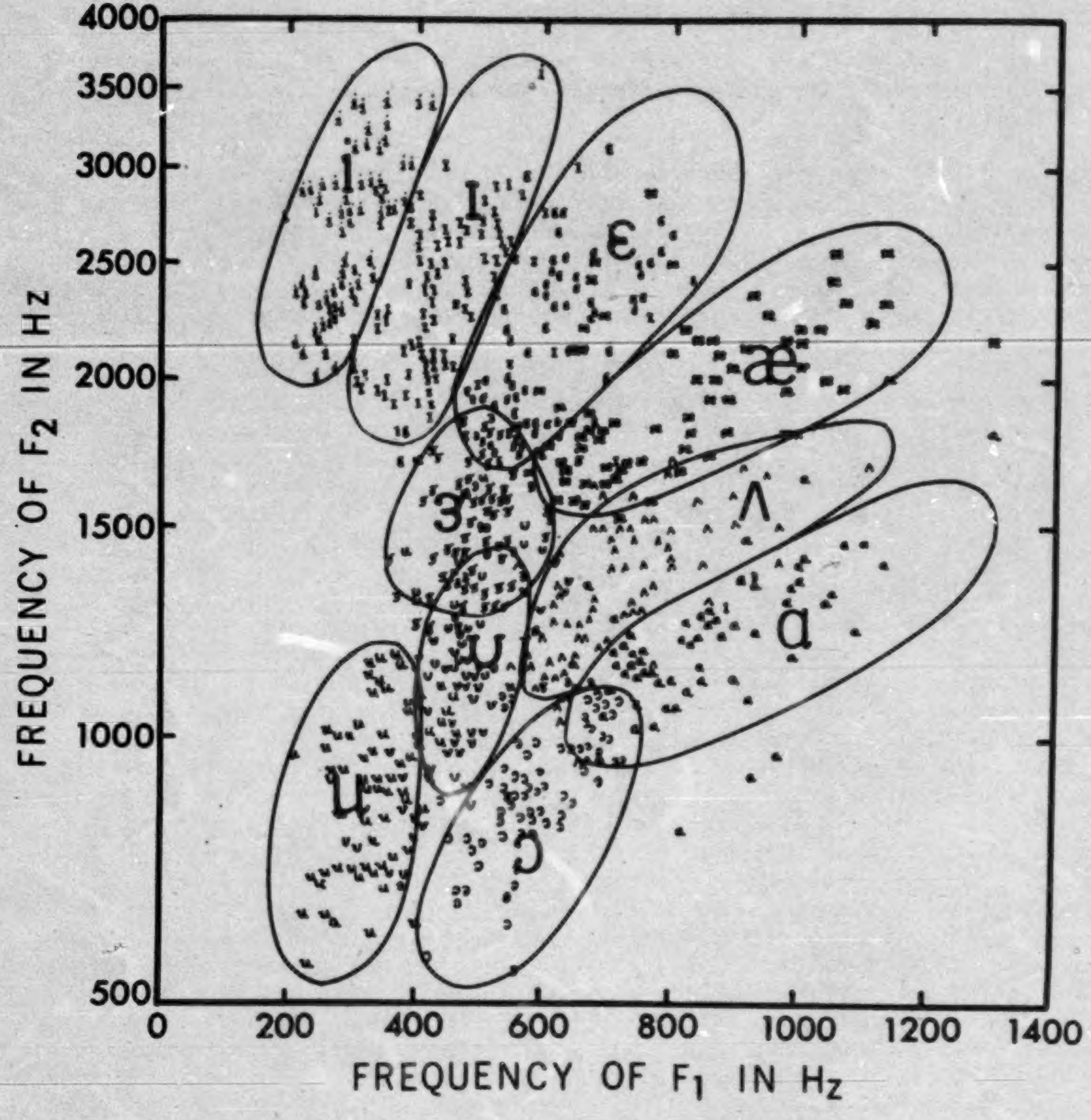

Figure 2: First- and second-formant frequencies of American-English vowels for a sample of 76 adult men, adult women, and children. The closed loops enclose 90 percent of the data points for each vowel category (redrawn from Peterson and Barney, 1952).

different spectra are heard by listeners as the same vowel, on what does the judgment of sameness depend? It cannot, he concludes, be due to any evidence in the sound:

> Therefore the identification is based on outside evidence....If this (outside evidence) were merely the memory of what the same phoneme sounded like a little earlier in the conversation, the task of interpreting rapid speech would presumably be vastly more difficult than it is. What seems to happen, rather, is this. On first meeting a person, the listener hears a few vowel phones and on the basis of this small but apparently sufficient evidence he swiftly constructs a fairly complete vowel pattern to serve as background (coordinate system) upon which he correctly locates new phones as fast as he hears them....(p. 61).

Thus, in Joos's view, the listener calibrates the talker's vowel space on the basis of a small subset of sample utterances. The listener needs some reference points to define the range and distribution of the talker's vowels. These reference signals, Joos suggests, could be supplied by extreme articulations (in terms of tongue height and point of tongue contact). Thus, Joos (1948) leaves no doubt that the coordinate system he has in mind is based at least in part on a model of the vocal tract:

> The process of correctly identifying heard vowel colors doubtless in some way involves articulation. A person who is listening to the sounds of a language that he can speak is not restricted to merely acoustic evidence for the interpretation of what he hears, but can and probably does profit from everything he "knows," including of course his own way of articulating the same phones.

Since the publication of Joos's work, a normalization step in speech perception has been assumed by virtually everyone who has written on the subject, whether or not the writer accepted Joos's version of the motor theory of speech perception (or, indeed, any other version of motor theory). The idea that a listener makes use of reference vowels for calibration of a talker's vowel space has also persisted. Gerstman (1968) and Lieberman (1973)--whose contributions we shall discuss presently--each have taken up the reference vowel idea and defended it. What is surprising, given this overwhelming consensus, is how few attempts have been made to measure the ambiguity in perception of vowels that is directly attributable to talker variation.

Joos's (1948) statement of the constancy problem implies that, for an unknown talker, isolated syllables should be highly ambiguous from the standpoint of the perceiver. It is perplexing, then, that two experiments that directly measured the perceptual ambiguity of natural speech found little support for this prediction. Peterson and Barney (1952), to whom we are indebted for systematic acoustic measurements of individual differences in vowel formants, also attempted to assess the perceptual consequences of the variation they discovered in production. The same recorded utterances used in making the spectrographic measurements were also assembled into listening tests. Listeners had to identify tokens of 10 vowels in /h-d/ consonantal environment; the set consisted of heed, hid, head, had, hod, hawed, hood, who'd, hud, and heard. Syllables produced by groups of 10 talkers (men, women, and children) were randomly mixed on each listening test, ensuring that opportunities for normalization would be slight. Although plots of the first two formants show that the regions occupied by these vowels overlap considerably, perceptual judgments were remarkably accurate, with 94 percent of the words perceived correctly. A similarly low error rate for perception of a larger set of vowels, which included diphthongs, was reported by Abramson and Cooper (1959).

These perceptual data do not support the notion that a single syllable, spoken in isolation, is necessarily ambiguous. Apparently, the information contained within a single syllable is usually sufficient to allow whatever adjustment for individual talker characteristics might be required. The history of the research following Peterson and Barney's (1952) study shows that this conclusion was not generally drawn.

The work of Ladefoged and Broadbent (1957) and Ladefoged (1967) is widely cited as evidence that the listener has relative criteria for vowel identification and that the identity of a vowel depends on the relationship between the formant frequencies for that vowel and the formant frequencies of other vowels produced by the same talker. As a result, vowel space must presumably be rescaled for each voice a listener encounters. The Ladefoged and Broadbent (1957) study was designed to find out whether subjects could be influenced in their identifications of a test word by variations in an introductory sentence preceding it. Synthetic speech was used in order to gain precise control over the acoustic parameters. A set of test syllables of the form /b-vowel-t/ was prepared on the synthesizer. Listening tests were made up in which the test words were presented following a standard sentence: Please say what this word is. Variants of this sentence were produced by shifting the frequencies of the first or second formants up or down. Each was intelligible despite wide acoustic differences. The results showed that the same test word was identified as bit when preceded by one version of the test sentence (i.e., one "voice") and as bet when preceded by a second version (another "voice") in which the first formant varied over a lower range. The authors conclude from perceptual shifts such as this that the identification of a vowel is determined by its relation to a prior sample of the talker's speech (provided here by the test sentence).

Ladefoged (1967) interprets these findings within the framework of Helson's (1948) adaptation level theory. This theory attempts to account for the extraordinary efficiency of the compensatory mechanisms that achieve constancies, such as color constancy under changing illumination, by supposing that the perceiver scales his responses not to the absolute properties of each stimulus, but according to the weighted mean of a set of stimuli distributed over time. The introductory sentence, in the Ladefoged and Broadbent experiment (1957), is understood as providing the standard or anchor, thus creating an internal adaptation level to which the test words are referred. We shall return to the adaptation level hypothesis presently, after we have introduced some relevant findings of our own.

If it is true that a listener needs a sample of speech in order to fix the coordinates of a talker's vowel system, we need to know how large a sample is required and whether particular vowels are more effective than others as "anchors." As we noted earlier, Joos (1948) believed that the best reference signal would be one that allows the listener to determine the major dimensions of the talker's vocal tract. He therefore suggested that the "point vowels"

/i/, /ɑ/, and /u/ might be the primary calibrators of vowel space, since they are the vowels associated with the extremes of articulation. Lieberman and colleagues (Lieberman, Crelin, and Klatt, 1972; Lieberman, 1973) agree that these vowels probably play an important role in disambiguating syllables produced by a novel talker. They note that the point vowels are exceptional in several ways: they represent extremes in acoustic and articulatory vowel space, they are acoustically stable for small changes in articulation (Stevens, 1972), and they are the only vowels in which an acoustic pattern can be related to a unique vocal-tract area function (Lindblom and Sundberg, 1969; Stevens, 1972).

Gerstman (1968) has made one of the few direct attempts to test the idea that a subset of vowels can serve to calibrate a talker's vowel space and reduce errors in recognition of subsequently occurring vowels. Gerstman developed a computer algorithm that correctly classified an average of 97 percent of the vowels in the syllables produced by the Peterson and Barney panel of 76 talkers. For each talker's set of 10 utterances, the program rescaled the first- and second-formant frequency values of each medial vowel, taking the extreme values in the set as the endpoints. Since these extreme formant values are typically associated with /i/, /ɑ/, and /u/, the procedure corresponds to a normalization of each talker's vowel space with reference to his own utterances of the point vowels. The classification system was essentially a filter bank that classified the vowels according to the scaled values for the first two formants, and the sums and differences of these values. By inserting the normalization stage between segmentation and classification, Gerstman's program succeeded in reducing by half the errors in classification made by human perceivers [recall that Peterson and Barney's (1952) listeners made 6 percent errors]. We must keep in mind however, that a successful algorithm is not a perceptual strategy, but only a possible strategy. Although it is of interest that such an algorithm can, in principle, serve as the basis for categorization of the signals, we are aware of no evidence that the human perceptual apparatus functions analogously. For example, it would be necessary to demonstrate that humans scale individual formants and calculate sums and differences between them.

It seemed to us that speculation had far outstripped the data bearing on the vowel constancy problem. In fact, the few studies of perceivers' recognition of natural (as distinguished from synthetic) speech indicated that isolated syllables spoken by novel talkers are remarkably intelligible. Therefore, it seemed important to verify Ladefoged and Broadbent's (1957) demonstration with natural speech in which all the potential sources of information that are ordinarily available to perceivers are present in the signal. Similarly, Gerstman's (1968) success in machine recognition using the three point vowels as calibrating signals needed to be evaluated against the performance of human listeners. Accordingly, we designed experiments to determine the size of the perceptual problem posed for the listener when the speaker is unknown. This involved a comparison of perceptual errors, under matched conditions, when the test words were spoken by many talkers and when all were produced by only one talker. We also sought to discover whether certain vowels (e.g., the point vowels) have a special role in specifying the coordinates of a given talker's vowel space.

HOW DOES A LISTENER MAP A TALKER'S VOWEL SPACE? AN EXPERIMENT TO DETERMINE THE SIZE OF THE PROBLEM

We (Verbrugge, Strange, and Shankweiler, 1974) first attempted to measure the degree of ambiguity in vowel perception attributable to lack of congruence of the vowel spaces of different talkers. To measure this, we presented listeners with unrelated words or nonsense syllables, so that broad linguistic context would make no contribution to the act of identification. Our studies of this problem were fashioned after Peterson and Barney (1952). We presented nine vowels in the consonantal environment /p-p/; thus, the set consisted of /pip/, /pɪp/, /pɛp/, /pæp/, /pɑp/, /pɔp/, /pʌp/, /pʊp/, and /pup/. In one listening test, the listeners heard tokens spoken by 15 different voices (5 men, 5 women, 5 children), arranged in random order. To determine what proportion of perceptual error is due to uncertainty about the talker, as opposed to other sources, a second listening test was employed in which a single talker uttered all the tokens on a given test. The talkers were given no special training. They were urged to recite the syllables briskly in order to bring about some undershoot of steady-state targets. Our objective was to achieve conditions as similar as possible to normal conversational speech.

Listeners misidentified[6] an average of 17 percent of the vowels when spoken by the panel of 15 talkers, and 9 percent when each of three tests was spoken by a single individual (a representative man, woman, and child from their respective groups). The difference between these two averages, 8 percent, is a measure of the error attributable to talker variation. Although this is a statistically significant difference [$\underline{t}(50\ df) = 5.14$, $\underline{p} < .01$], its absolute magnitude is surprisingly small. Less than half the total number of errors obtained in the variable-talker condition can be attributed to talker variation.

Listeners can identify vowels in consonant-vowel-consonant syllables with considerable accuracy even when they are spoken by an assortment of talkers deliberately chosen for vocal diversity. The intended vowel was identified on 83 percent of the tokens in a test designed to maximize ambiguity contributed by vocal-tract variation.[7] In a second study, listeners were asked to identify 15 vowels (monophthongs and diphthongs) spoken in /h-d/ context by 30 talkers. Here, the rate of identification errors was 13 percent overall and 17 percent for the nine monophthongs alone. The results of both studies are in essential agreement with earlier perceptual data reported by Peterson and Barney (1952) and Abramson and Cooper (1959), although the error rates obtained by these investigators were even lower than those we obtained.[8]

---

[6] We have taken the intended vowel of the talker as the criterion of correct identification. That is, we have defined an identification error as a response by the listener that does not correspond to the phonemic category intended by the talker. It might be the case that errors so defined are as much due to mispronunciation as to misperception. No correction was given to talkers during recording other than to clarify orthographic confusions in a few instances. In the case of the youngest children, some coaching was required before they pronounced the nonsense syllables. However, no adult models were provided immediately prior to utterances that were included in the tests.

[7] Each of the 15 talkers spoke only three tokens containing different vowels. These tokens were separated in the test by no fewer than eight intervening tokens spoken by different talkers. Listeners were unable to judge how many talkers were included on a test.

These findings do not bear out the common assumption of a critical need for extended prior exposure to a talker's speech. The information contained within a single syllable appears to be sufficient in most cases to permit recognition of the intended vowel; familiarity with a voice seems to play a rather small role in the identification process.

ARE THE POINT VOWELS USED BY LISTENERS AS AIDS TO NORMALIZATION?

We (Verbrugge, Strange, and Shankweiler, 1974) next examined the possibility that an introductory set of syllables increased the likelihood that a succeeding vowel produced by the same talker was correctly recognized. Because we wished to test a specific hypothesis about the stimulus information required for normalization, we did not employ an introductory sentence as Ladefoged and Broadbent (1957) had done. Instead, we introduced each target syllable by three precursor syllables; this provided three samples of the talker's vowels and little else. In one condition of the experiment, the precursors were /hi/, /hɑ/, and /hu/; these syllables contain examples of the talker's point vowels. For a second condition, we chose /hɪ, hæ , hʌ/, a set of nonpoint vowels that (like the point vowels) are quite widely separated in the space defined by the first and second formants.

Neither set of precursor syllables brought about a systematic reduction of perceptual errors in identifying the target vowels. The errors in each precursor condition averaged 15 percent (compared to 17 percent in the earlier condition without precursors), but in neither case does this difference approach significance.[9] The principal effect of the /hi/, /hɑ/, /hu/ precursors was to shift the pattern of responses somewhat, some vowels showing improved identification, others showing poorer identification.

The idea that normalization is specifically aided by the point vowels--as suggested by Joos (1948), Gerstman (1968), and Lieberman (1973)--is not supported by these data. In fact, no precursor syllables that we tried were found to have a systematic effect. A single, isolated syllable is usually sufficient to specify a vowel; prior exposure to specific subsets of vowels could not be shown to supply additional information. It would seem unnecessary to invoke a psychophysical weighting function in order to establish an internal adaptation level (cf. Ladefoged, 1967). We may surmise that the isolated syllable is not so ambiguous an entity as is sometimes implied.

---

[8] We suspect that these studies made somewhat less severe demands on listeners' perceptual capacities than our own. In the Peterson and Barney (1952) study, listeners heard only 10 different talkers on a particular test. Each talker spoke two tokens of each of 10 /h-d/ syllables. The study yielded an overall error rate of 6 percent. The Abramson and Cooper (1959) study employed eight talkers, each of whom spoke one token of 15 /h-d/ syllables. The overall error rate in that study ranged from 4 to 6 percent. An additional source of perceptual difficulty in our tests is the fact that /ɑ/ and /ɔ/ are homophonous in the dialect of most of our talkers.

[9] This result was confirmed in a separate study (cf. Verbrugge, Strange, and Shankweiler, 1974) of 15 vowels in /h-d/ context. When no precursors were present, errors averaged 13 percent. When each /h-d/ syllable was preceded by the syllables, /kip/, /kɑp/, /kup/, 12 percent of the responses were errors. The difference was not significant.

Because of the repetitive and stereotyped manner in which the precursors were presented, some readers might be inclined to doubt whether listeners made full use of the phonetic information potentially available and therefore to question whether these experiments are adequate to test the hypothesis. We can reply to this objection indirectly by referring to a further experiment in which the same precursor syllables did produce a measurable effect on perception of a subsequent target syllable. This experiment involved the same 15 talkers as the earlier experiment, but differed in that the test syllables were produced in a fixed sentence frame: *The little /p-p/'s chair is red*. Each talker was instructed to produce the sentence rapidly, placing heavy stress on the word *chair*. The unstressed, rapidly articulated /p-p/ syllables were excised from the tape recording and assembled into two new listening tests. In one condition, the /p-p/ targets were prefaced by the same tokens of the /hi, hɑ, hu/ precursors employed in the previous experiment. On this test, listeners made an average of 29 percent errors in identifying the vowels in the target syllables. In the other condition, no precursors were present. On this test, listeners misidentified 24 percent of the same vowels. Thus, misperception of target vowels occurred with significantly *greater* frequency when they were preceded by precursors [t(35 df) = 2.88, $p < .01$]. We may suppose that the precursors impaired recognition of succeeding vowels in this instance because they specified a speaking rate slower than that at which the /p-p/ syllables were actually produced. Thus, whereas we failed to find evidence for effects of precursors on normalization of vocal-tract differences, we do find evidence for adjustment to a talker's tempo (as hypothesized by Lindblom and Studdert-Kennedy, 1967), on the basis of preceding segments of speech.

THE ROLE OF FORMANT TRANSITIONS IN VOWEL PERCEPTION

Our results suggest that the identity of a vowel in a syllable spoken by a new talker is likely to be specified by information within the syllable itself. The phonetic context supplied by preceding syllables apparently serves a function other than that of adjustment for a new set of vocal-tract parameters: it may enable the perceptual system to gauge the tempo of incoming speech and to set its criteria accordingly. We were encouraged by these preliminary findings to look for the sources of information that specify the vowel within the syllable, and to explore how that information is used by the perceiver in the process of vowel perception.

As we noted earlier, the formant transitions in a syllable vary systematically as a function of both the consonant and the vowel. Therefore, we might expect that the listener utilizes information contained in the transitions in recovering the identity of the medial vowel. Research on the identification of isolated steady-state vowels (i.e., vowels that are *not* coarticulated with consonants) indirectly supports this expectation. Perception of isolated vowels is notably unreliable. Fairbanks and Grubb (1961) presented nine isolated vowels produced by seven phonetically trained talkers to eight experienced listeners. The overall identification rate was only 74 percent; rates for individual vowels ranged from 53 to 92 percent. Slightly better identification of isolated vowels was obtained by Lehiste and Meltzer (1973) for three talkers, where, again, talkers and listeners were phonetically skilled. Fujimura and Ochiai (1963) directly compared the identifiability of vowels in consonantal context and in isolation. They found that the center portions of vowels, which had been gated out of CVC syllables, were less intelligible in isolation than in syllabic context.

Research bearing on this question has also been done with synthetic speech. Millar and Ainsworth (1972) reported that synthetically generated vowels were more reliably identified when embedded in an /h-d/ environment than when acoustically identical steady-state target values were presented in isolation. Finally, Lindblom and Studdert-Kennedy (1967) noted that listeners used different acoustic criteria to distinguish pairs of vowels depending on whether judgments were made on isolated vowels or on the same vowel targets embedded within a CVC environment.

There are at least two ways that the transitional portions of the acoustic signal might provide information for vowel identity. One possibility is that transitions play a role in specifying talker characteristics. Since the loci of formant transitions for a particular consonant vary with differences in vocal-tract dimensions (Fourcin, 1968; Rand, 1971), transitions might serve as calibration signals for normalization. Particularly when the phonemic identity of the consonants is fixed and known to the listener, the transitions might serve to reduce the ambiguity of the vowel by providing information about vocal-tract characteristics of the talker who produced the syllable.

We may also envision a second possibility that is at once more general and more parsimonious: the acoustic specification of vowels, like consonants, is carried in the dynamic configuration of the syllable. In other words, the syllable as a whole cospecifies both consonants and vowel. In this view, transitions may be regarded as belonging to the vowel no less than to the consonants. If this were true, we would expect that the perception of medial vowels would be aided by the presence of consonantal transitions regardless of whether the perceiver encounters many talkers on successive tokens or only one.

To make an experimental test of these possibilities, we (Strange, Verbrugge, and Shankweiler, 1974) constructed a new set of listening tests that contained a series of isolated vowels. In one condition the vowels were spoken by the same panel of 15 talkers described above. In a second condition, a single talker produced the full series of vowels. Together with the earlier tests with /p-p/ syllables, these materials allowed us to compare the relative effects on vowel identifiability of two major variables: presence or absence of consonantal environment and presence or absence of talker variation within a test. This also placed us in a position to evaluate the alternative hypotheses about how consonantal environment contributes to vowel perception.

According to either hypothesis, we would expect that the perception of isolated vowels would be less accurate than the perception of medial vowels on a listening test in which the tokens are produced by different talkers. However, the two alternative hypotheses generate different expectations concerning the error rate on isolated vowels and medial vowels when the talker does not vary within a test. If the advantage of consonantal environment is due to use of transition cues for normalization, we could expect to obtain no difference between performance on these two conditions, because in neither case is there a need for repeated calibration. Therefore, we would expect that vowel recognition would be as accurate for the isolated vowels as for the medial vowels. If, on the other hand, the consonantal environment provides critical information for the vowel independent of talker-related variation, we would expect a difference in consonantal environment to affect performance whether or not talkers vary within a test. Thus, we would expect identification of isolated vowels to be less accurate than medial vowels even for tests in which the talker did not vary.

The results for the isolated vowel tests support the latter hypothesis. The average error in the variable-talker condition was 42 percent (compared to 17 percent errors on the comparable test in which vowels were spoken in /p-p/ environment). This increase in errors is consistent with either hypothesis. However, the results for the single-talker condition also showed a large increase in errors when there was no consonantal environment. The average error in the single-talker conditions was 31 percent for the isolated vowels (compared to 9 percent errors on medial vowels). Moreover, a vowel-by-vowel comparison showed that for every vowel in both talker conditions, the error rate on the isolated vowel was greater than on the corresponding medial vowel. Both major variables (Consonants Present versus Absent and Talker Variation Present versus Absent) were shown to produce significant differences in overall errors [$F(1,94\ df) = 125.17$ and 21.18, respectively, $p < .01$]. The decrease in accuracy of vowel recognition due to the absence of consonantal environment was approximately the same whether talkers varied or not (i.e., the analysis showed no significant interaction between variables). We may surmise, therefore, that consonantal transitions do not aid in specifying a vowel by providing information for a normalization stage. On the contrary, these results indicate that the presence or absence of transitions is much more critical for accurate recognition than the degree of experience with a talker's vocal-tract parameters. Whereas the presence of within-test talker variation impairs recognition by only about 8 percent, the absence of a consonantal environment impairs performance by more than 20 percent.

The possibility cannot be overlooked, however, that the relatively poor perception of isolated vowels is attributable primarily to the talkers' inability to produce them reliably. Since isolated vowels do not occur in natural speech (with a few exceptions), talkers may produce them in peculiar ways, with formant frequencies uncharacteristic of the values found in natural syllables. Also, the characteristic relative durations of the vowels (Peterson and Lehiste, 1960) might not be preserved by talkers in their productions of isolated vowels.

To investigate these possibilities, we undertook spectrographic analysis of the tokens of isolated vowels and medial vowels used in our listening tests. Center frequencies of the first three formants and vowel duration were measured for all the tokens in the variable-talker tests, as well as for tokens of all nine vowels spoken in isolation by each of the 15 talkers. The data provided no evidence that the isolated vowels were produced in an aberrant manner. Average formant frequencies for men, women, and children correspond quite closely to those reported by Peterson and Barney (1952) (for vowels in /h-d/ environment), with the exception of /ɔ/.[10] When the formant frequencies of each talker's isolated and medial vowels are compared, the values are found to be highly similar. Measurements of vowel duration also fail to account for the increased error rate for isolated vowels. Although the durations of these were for the most part longer than the vowels in /p-p/ environment, the relative durations of the nine isolated vowels were much the same as the relative durations of vowels in consonantal environment. We may suppose, therefore, that the higher error rate for isolated vowels compared to that for vowels in a fixed consonantal environment cannot be explained on the grounds that isolated vowels tend to be produced in an aberrant manner.

---

[10] This deviation is due to a dialect difference between our group of talkers (predominantly natives of the upper Midwest) and Peterson and Barney's group.

The message of these perceptual data is clear: isolated, sustained vowels, although they correspond well to the phonetician's idealized conception of a vowel,[11] are poorly specified targets from the standpoint of the listener. Lehiste and Peterson (1959) found that many hours of practice were needed by untrained listeners before they could identify isolated vowels accurately, even when the tokens were painstakingly produced by a single phonetically trained talker. The ability to identify these "ideal" vowels may be a highly specific skill with little relevance to the identification of vowels in natural speaking situations.

At this point the objection might still be raised that the tests used to measure the perceptual difficulty of medial vowels are unrepresentative of natural conditions. One possibility is that there may be an advantage associated with consonantal context if the context is known beforehand (/p-p/ in this case), but that this advantage would be largely eliminated if the identity of neighboring consonants were unknown (as is often the case in natural speech). To test this possibility, we constructed a listening test in which the target vowels were enclosed by a variable consonantal environment. A panel of 12 talkers (a subset of the original 15) spoke a series of consonant-vowel-consonant syllables. In each syllable, one of the six stop consonants (/b, d, g, p, t, k/) appeared before the vowel and one of the six appeared after the vowel; consonants were selected so that each occurred equally often in each position. One group of listeners was asked to identify only the vowel in each test token; a second group was asked to identify the two consonants as well as the vowel. The average error in identifying the vowels was 22 percent for the first group and 29 percent for the second. Both error rates are well below the 42 percent error rate obtained on the variable-talker test with isolated vowels. In other words, even when listeners do not know the identity of either the consonants or the vowel, recognition is significantly more accurate for medial vowels than for isolated vowels.

A second possible objection to the earlier tests with medial vowels might be that syllables spoken in isolation (in "citation form") are unrepresentative of the syllables found in rapid, connected speech. The medial vowels in rapidly spoken syllables might be at least as difficult to identify as isolated vowels, since the vocalic portions of such syllables often fail to reach the steady-state values characteristic of syllables spoken in citation form. The study reported earlier in this paper bears directly on this question. When /p-p/ syllables spoken in unstressed position are excised from a carrier sentence and assembled into a listening test, listeners made an average of 24 percent errors in perceiving the medial vowels. This is not much greater than the 17 percent error rate for perception of /p-p/ syllables read from a list, but is substantially less than the 42 percent error rate for isolated vowels. One might have guessed that the brevity of the short /p-p/ syllables and their failure to reach steady-state values would make them more difficult to identify than isolated vowels, which are longer in duration and more stable acoustically. Apparently, the presence of a consonantal environment more than compensates for these difficulties.

---

[11] The formant space is less compressed for isolated vowels than for medial vowels. Thus, if the values of static first and second formants were the primary carriers of vowel quality, isolated vowels should be better perceived because their acoustic values are more widely separated.

CONCLUDING REMARKS ON THE PROBLEM OF VOWEL CONSTANCY

Let us consider what we have learned about how the perceiver might achieve constancy of vowel quality. In our studies of vowel perception, the objective was to isolate sources of vocalic information in the natural speech signal. We employed signals that presented as many characteristics as possible of normal conversational speech, including a representative range of signal variations that result from physical differences among talkers.

Each way of conceptualizing the vowel contains an implied solution to the problem of perceptual constancy. We first considered the assumption that the vowel can be characterized by a steady-state output of the vocal tract, and that, to a first approximation, fixed-formant loci are associated with each vowel quality of all speakers. To the extent that this assumption is correct, the constancy problem is trivial. Only minor adjustments for variation would be required.

We saw that this conception of the vowel as a simple acoustic event, segmented in time and in spectral frequency composition, was widely shared among students of speech, including those who initiated earlier attempts at automatic speech recognition. We have reviewed a number of findings that are incompatible with this view. First, steady states are the exception, not the rule, in continuous natural speech. As a result of coarticulation of vowels with preceding and following consonants, the syllable is not discretely partitioned, and the information for the vowel is smeared throughout the syllable. Moreover, the variability occasioned by the phonemic environment of a vowel is compounded by the changes that accompany different speaking rates and different vocal-tract sizes. In retrospect, it is easy to see why attempts to design a generally useful speech recognition machine have so far failed.

A more sophisticated conception of the vowel acknowledges the problem of variability but continues to assume that vowels, even in running speech, can be perceived with reference to a single set of acoustic values. This view proposes that tokens of the "same" vowel fall on a line in vowel space defined by the first and second formants. The formant frequencies of two talkers' vowels would then be constant multiples of one another. We noted that this relationship could not literally hold, because vocal tracts differ in shape as well as in size. This rules out an analogy to a melody played in a different key, or to a magnetic tape recording played back at a different speed. The failure of these analogies is revealed by Peterson and Barney's (1952) measurements of first- and second-formant frequencies in men, women, and children. The results of the measurements (displayed in Figure 2) showed wide dispersion of formant values for different speakers with considerable overlap of formants for neighboring vowels. Even when one considers only those tokens on which perfect agreement was obtained by listeners, much scatter among formant values is observed (cf. Peterson and Barney's Figure 9).

Failing to find the invariant relation preserved by linear scaling, investigators have sought a transformation that might yield a closer approximation. For example, it has become an accepted practice to plot units of frequency (Hz) on a scale of mels (Peterson, 1961; Ladefoged, 1967).[12] Transformation of formant frequencies to mels might be defended on the grounds that this unit reflects the response of the auditory system to frequency. However, we are skeptical that the constancy problem can be illuminated by a search for the right scale factor. Ladefoged (1967), who attempted to reduce variability by employing phoneticians as talkers, concluded that separation of all vowels cannot be attained by scaling the first- and second-formant frequencies, whether in linear fashion or nonlinearly, as on a scale of mels.

Although no one has succeeded in demonstrating a generally applicable scaling (normalizing) function, it is widely assumed that perceivers must apply such a function to each new talker they encounter. There has been speculation about the minimal stretch of speech required for calibration. Ladefoged's (1967) application of adaptation-level theory to the problem of speaker normalization reflects the common assumption that some extended sample of a new talker's utterances is required for determining the weights that enable the normalizing adjustments to be made. As we noted, Joos (1948) and Lieberman (1973) proposed that ambiguity of a new talker's utterances can be resolved by reference vowels that permit the perceiver to construct a model of the talker's vowel space, scaling the input according to parameters derived from these calibration vowels.

Listeners can apparently adjust their criteria for perception of synthetic vowels according to the formant ranges specified by a precursor sentence (Ladefoged and Broadbent, 1957). The successful performance of Gerstman's (1968) normalizing algorithm indicates that frequencies of the first and second formants could, in principle, suffice for this purpose. However, we doubt, as does Ladefoged (1967) himself, that first- and second-formant frequencies exhaust the sources of information that specify a vowel in the natural speech signal.[13] Moreover, the fact that listeners can perceive randomly ordered syllables accurately, indicates that there is little need for a mechanism that requires a sample of several syllables in order to construct a normalization schema. Finally, in our own experiments with natural speech, we failed to find that point vowel precursors, or another set of widely spaced vowels, brought about a systematic improvement in recognition of the following vowel.

Our results do not, therefore, support the view that vowels are relational values in a metric space that must be scaled according to other vowels produced by the same talker. If that theory were correct, it is difficult to see how precursors could fail to improve recognition of an immediately following medial vowel. The presence of coarticulated consonants within the syllable proved far more useful for categorizing natural vowels than prior experience with a talker's utterance. In sum, our studies failed to provide supporting evidence for current conceptions of the normalization process. They force us to consider whether there is justification for a separate and preliminary normalization stage in speech perception.[14]

---

[12] A mel is a psychophysical unit reflecting equal sense distances of pitch and bearing an approximately logarithmic relation to frequency for frequencies above 1000 Hz (Stevens and Volkmann, 1940).

[13] This is also Peterson's (1961) conclusion, based on studies of filtering.

A major difficulty with all the proposals stated above is that they view the invariance problem in terms of the relation among formant frequencies of relatively sustained vowels. Even if such efforts to discover an algorithm were successful, they would not be sufficient to explain perception of vowels in natural conversational speech, because in such utterances a region of steady-state energy is rarely present in the signal. The presence of sustained acoustic energy at certain "target" frequencies is not essential for identification of a vowel under natural listening conditions. On the contrary, there is evidence that changing spectral patterns are much superior to sustained values as carriers of vowel quality. We cited earlier reports that isolated steady-state vowels are poorly perceived, even after listeners were given substantial training and when the target vowels were spoken by phonetically trained talkers (Fairbanks and Grubb, 1961; Lehiste and Meltzer, 1973). The results of our perceptual studies definitely confirm the perceptual difficulty of isolated vowels. Listeners misidentified 31 percent even when all items within a given test list were produced by the same talker. Moreover, vowels coarticulated with surrounding consonants, as is normal in running speech, were considerably more intelligible than isolated vowels spoken by the same talkers (e.g., 9 percent of vowels in /p-p/ environment were misidentified). It seems unlikely, therefore, that the perceptual system operates by throwing away information contained in formant transitions. Indeed, Lindblom and Studdert-Kennedy (1967), in studies with synthetic speech, demonstrated that listeners use this information directly in their placement of vowel phoneme boundaries. Vowel identifications varied with direction and rate of transitions even when the formant frequency values at the syllable centers were held constant. In short, it is futile to seek a solution to the constancy problem by analysis of any acoustic cross section taken at a single instant in time, and we must conclude that the vowel in natural speech is inescapably a dynamic entity.

Heuristic procedures for automatic recognition of consonants often begin by guessing the identity of the coarticulated vowel, since it is known that the specific shapes of the formant transitions are conditioned by the vowel. The vowel is assumed to be a stable reference point against which the identity of the consonant may be determined. But this, of course, presupposes that the vowel is more directly available than the consonant. We have now examined a number of indications that the problem of perceptual constancy may be no less abstract for the vowel than for the consonant. We found that isolated steady-state vowels provided especially poor perceptual targets. Moreover, there were not substantially more errors in identifying the medial vowels of rapidly spoken syllables (where steady states were presumably not attained), compared to errors on medial vowels in syllables read from a list. It is clear that we can no longer think of defining vowels in terms of acoustic energy at characteristic target frequencies. A recognition device based on filters tuned to specific frequencies is as unworkable for vowels as it is for consonants. If the idea of the vowel target is to be retained, it must take account of the dynamic character of the syllable.

---

[14] But see Summerfield and Haggard (1973). These investigators measured an increase in reaction time to synthetic syllables from different (simulated) vocal tracts, which they interpret as reflecting extra processing time required for a normalization stage.

Lindblom's (1963) conception has this virtue. Formant contours are characterized as exponential functions that tend toward asymptotic "target" values associated with the vowel nucleus. Thus, a target can be defined acoustically even though it corresponds to no spectral cross section through the syllable. Our perceptual data presented here are compatible with this view that vowels are specified by contours of moving formants with certain invariant properties over stretches of approximately the length of a syllable.[15]

We may find a parallel to this conclusion in studies of speech production. Investigation of the manner in which phonemes are joined in the syllable reveals context-dependent relationships similar to those we have noted in the acoustic signal. Lindblom's (1963) dynamic theory of vowel articulation is, in fact, an attempt to explain how contextual influences on the acoustic and phonetic properties of vowels are produced. According to this view, undershoot in running speech is brought about by inertia in the response of the articulators to motor excitations occurring in rapid temporal succession. Invariant neural events corresponding to vowel targets thus fail to bring articulators to the positions they assume when the vowel is produced in a sustained manner.

Lindblom's (1963) inference of articulatory undershoot during rapid speech has been confirmed by cinefluorographic data (Gay, 1974). His account of the mechanism of undershoot has not gone unchallenged, however. MacNeilage (1970) presented a different view of the inherent variability of speech production. Basing his conclusions on extensive electromyographic studies of context effects in articulation, he argued that variability of muscle contraction is not to be understood merely as an unfortunate consequence of mechanical constraints on articulator motion, but as necessarily built into the system in order to permit attainment of relatively invariant target shapes. Gay (1974) cited his cinefluorographic findings in support of MacNeilage's hypothesis that variability of gesture must be regarded as a design characteristic. However, unlike MacNeilage, he concluded that there was variability not only of gesture, but of spatial target, since he failed to find invariance of vocal-tract shape for a central vowel /ɑ/ that held across speaking rates and consonantal environments. The same conclusion was drawn by Nooteboom (1970), who argued that the kinds of reorganization that occur in talking with the teeth clenched make it difficult to retain the idea of invariant spatial targets. Just as the attainment of a specific acoustic target value is not necessary for the successful perception of a vowel, it is probably the case that the attainment of a specific target shape is not necessary for its effective production. Thus, in production as in perception, it has become increasingly difficult to entertain the notion of an invariant target for each vowel, as long as the meaning of invariance is restricted to a specific vocal-tract shape and its resonances. It is likely that the units of production, like the units of perception, cannot be defined independently of the temporal dimension. Speech, viewed either as motor gesture or as acoustic signal, is not a succession of static states. Invariance in vowel production, then, can be discovered only in the context of the dynamic configuration of the syllable.

---

[15] Our understanding of the vowel has been influenced by Gibson's (1966) approach to the problems of event constancy in visual perception, which is to seek regularities in the stimulus pattern that can only be defined over time. A similar approach is taken by Shaw, McIntyre, and Mace (1974), and by Shaw and Pittenger (in press). We tend to agree with these authors that the dynamic invariants specifying an event may be perceived directly by perceptual systems that are appropriately tuned. While Lindblom (1963) and Lindblom and Studdert-Kennedy (1967) offer a dynamic characterization of vowels, they apparently made the usual assumption that only temporal cross sections can be directly perceived, and they supposed that vowel perception is mediated by a process of analysis-by-synthesis in which the dynamic invariants are used to compute possible input patterns.

The reader might wonder at this point whether the various productions and acoustic forms of a vowel are so heterogeneous that no coherent physical definition (however abstract) could be found that embraces them all. Perhaps the required invariance is not to be found in the acoustic signal at all. If it is not, a radical solution to the constancy problem is to suppose that the variants of a phoneme are physically unrelated, and to assume that the brain stores separately a prototype of each vowel and consonant for every phonemic environment. If we can extend to speech perception an argument made by Wickelgren (1969) concerning its production, then Wickelgren's hypothesis of "context-sensitive allophones" is such a proposal. However, the proposal has little to recommend it. Halwes and Jenkins (1971) find a number of flaws, two of which are critical. First, the proposal fails to capture the phonological relations that are known to be important in understanding both the production and perception of speech. Second, it ignores the "creativity" inherent in the production of speech that permits the reorganization of articulatory movements to maintain intelligibility even when normal speech movements are blocked, as when talking with the teeth clenched or with food in the mouth, or when under the influence of oral anesthesia.

In light of the evidence we have surveyed, it is obvious that attempts to understand the psychophysical constancy relations in speech have failed to discover transparent isomorphisms between signal and perception. This failure has led many to doubt whether a psychophysics of speech could ever illuminate the constancy problem. But certainly, it does not follow from the complexity of the psychophysical relation that the signal fails to specify the phonemic message uniquely.[16] The emphasis that current theories place on the relational nature

---

[16] We doubt that a "distinctive feature" description of speech would allow a simpler psychophysical relation to be stated. Phonemes are often characterized by a set of component features that are the basis for contrastive phoneme pairs. For example, /b/ and /d/ contrast in place of articulation, while /d/ and /t/ contrast in voicing. There is substantial evidence that such features are integral to the perceptual analysis of speech. We believe that the same arguments apply to the detection of distinctive features as apply to the

of the vowel is misleading because it underestimates the richness of the signal in natural speech, a richness that is attested by the great tolerance of the perceptual system for a degraded speech signal (as in noisy environments or after filtering). To abandon the search for acoustic invariants because the psychophysical relations are complex would surely be a backward step. It should be appreciated, however, that commitment to the principle of invariance does not bind us to a literal isomorphism between signal and percept. The weight of evidence conclusively opposes a one-to-one mapping of perceptual segments and their dimensions on physical segments and their dimensions. In the case of vowels, we have argued that the invariants cannot be found in a temporal cross section but can only be specified over time.[17] For vowels, as for other phonological segments, a major goal of research is to discover the appropriate time domains over which invariance might be found.

REFERENCES

Abramson, A. S. and F. S. Cooper. (1959) Perception of American English vowels in terms of a reference system. Haskins Laboratories Quarterly Progress Report QPR-32, Appendix 1.

Allport, F. H. (1924) Social Psychology. (Boston: Houghton Mifflin).

Bloomfield, L. (1933) Language. (New York: Henry Holt).

Cole, R. A. and B. Scott. (1974a) The phantom in the phoneme: Invariant cues for stop consonants. Percept. Psychophys. 15, 101-107.

Cole, R. A. and B. Scott. (1974b) Toward a theory of speech perception. Psychol. Rev. 81, 348-374.

Cooper, F. S., P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman. (1952) Some experiments on the perception of synthetic speech sounds. J. Acoust. Soc. Amer. 24, 597-606.

Cooper, F. S., A. M. Liberman, and J. M. Borst. (1951) The interconversion of audible and visible patterns as a basis for research in the perception of speech. In Proceedings of the National Academy of Sciences 37, 318-328.

---

perception of phonemes; namely, they are specified abstractly over stretches of speech of varying length. Thus, a recognition model that includes feature recognition as an early stage must meet the same tests that we have outlined for phoneme recognition in general.

[17] A derived invariant, defined in terms of relations, has been described for the voicing distinction in consonants. In spectrographic analyses of voicing in stop consonants in many languages, Lisker and Abramson (1971) discovered a unity among the apparently diverse and unrelated acoustic features that are correlates of the voiced-voiceless distinction. Their work suggests that aspiration, explosion energy accompanying stop release, and first-formant intensity may all be understood in terms of control of the time relations between stop-closure release and the onset of laryngeal vibration. The derived cue, voice onset time, is a relatively invariant property of the signal for a given overall speaking rate. However, voice onset time is not a simple property in that it is definable only in terms of a temporal relation between two events occurring within the syllable. See Lisker (1975) and MacNeilage (1972) for discussions bearing on the importance of timing in speech.

Delattre, P. C., A. M. Liberman, F. S. Cooper, and L. J. Gerstman. (1952) An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesized from spectrographic patterns. *Word 8*, 195-210.

Denes, P. B. (1963) On the statistics of spoken English. *J. Acoust. Soc. Amer. 35*, 892-904.

Fairbanks, G. and P. A. Grubb. (1961) A psychophysical investigation of vowel formants. *J. Speech Hearing Res. 4*, 203-219.

Fant, C. G. M. (1960) *Acoustic Theory of Speech Production*. (The Hague: Mouton).

Fant, C. G. M. (1962) Descriptive analysis of the acoustic aspects of speech. *Logos 5*, 3-17.

Fant, C. G. M. (1966) A note on vocal-tract size factors and nonuniform F-pattern scalings. *Quarterly Progress and Status Report* (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) *QPSR-4*, 22-30.

Fant, C. G. M. (1970) Automatic recognition and speech research. *Quarterly Progress and Status Report* (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) *QPSR-4*, 16-31.

Fourcin, A. J. (1968) Speech source inference. *IEEE Trans. Audio Electroacoust. AU-16*, 65-67.

Fujimura, O. and K. Ochiai. (1963) Vowel identification and phonetic contexts. *J. Acoust. Soc. Amer. 35*, 1889(A).

Gay, T. (1974) A cinefluorographic study of vowel production. *J. Phonetics 2*, 255-266.

Gerstman, L. H. (1968) Classification of self-normalized vowels. *IEEE Trans. Audio Electroacoust. AU-16*, 78-80.

Gibson, J. J. (1966) *The Senses Considered as Perceptual Systems*. (Boston: Houghton Mifflin).

Halwes, T. and J. J. Jenkins. (1971) Problem of serial order in behavior is not resolved by context-sensitive associative memory models. *Psychol. Rev. 78*, 122-129.

Harris, C. M. (1953) A study of the building blocks in speech. *J. Acoust. Soc. Amer. 25*, 962-969.

Helson, H. (1948) Adaptation level as a basis for a quantitative theory of frames of reference. *Psychol. Rev. 55*, 297-313.

Hockett, C. F. (1958) *A Course in Modern Linguistics*. (New York: MacMillan).

House, A. S. and G. Fairbanks. (1953) The influence of consonant environment upon the secondary acoustical characteristics of vowels. *J. Acoust. Soc. Amer. 25*, 105-113.

Hyde, S. R. (1972) Automatic speech recognition: A critical survey and discussion of the literature. In *Human Communication: A Unified View*, ed. by E. E. David, Jr., and P. B. Denes. (New York: McGraw-Hill).

Joos, M. A. (1948) Acoustic phonetics. *Language*, *Suppl. 24*, 1-136.

Koenig, W. H., H. K. Dunn, and L. Y. Lacey. (1946) The sound spectrograph. *J. Acoust. Soc. Amer. 18*, 19-49.

Kuhl, P. (1974) Acoustic invariance for stop consonants. *J. Acoust. Soc. Amer.*, *Suppl. 55*, 55(A).

Kuhn, G. M. and R. McI. McGuire. (1974) Results of a spectrogram reading experiment. *Haskins Laboratories Status Report on Speech Research SR-39/40*, 67-79.

Ladefoged, P. (1967) *Three Areas of Experimental Phonetics*. (New York: Oxford University Press).

Ladefoged, P. and D. E. Broadbent. (1957) Information conveyed by vowels. J. Acoust. Soc. Amer. 29, 98-104.

Lehiste, I. and D. Meltzer. (1973) Vowel and speaker identification in natural and synthetic speech. Lang. Speech 16, 356-364.

Lehiste, I. and G. E. Peterson. (1959) The identification of filtered vowels. Phonetica 4, 161-177.

Liberman, A. M. (1957) Some results of research on speech perception. J. Acoust. Soc. Amer. 29, 117-123.

Liberman, A. M., F. S. Cooper, D. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.

Lieberman, P. (1973) On the evolution of language: A unified view. Cognition 2, 59-94.

Lieberman, P., E. S. Crelin, and D. H. Klatt. (1972) Phonetic ability and related anatomy of the newborn, adult human, Neanderthal man, and the chimpanzee. Amer. Anthropol. 74, 287-307.

Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. 35, 1773-1781.

Lindblom, B. E. F. and M. Studdert-Kennedy. (1967) On the role of formant transitions in vowel recognition. J. Acoust. Soc. Amer. 42, 830-843.

Lindblom, B. E. F. and J. Sundberg. (1969) A quantitative model of vowel production and the distinctive features of Swedish vowels. Quarterly Progress and Status Report (Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden) QPSR-1, 14-32.

Lisker, L. (1975) On time and timing in speech. In Current Trends in Linguistics, ed. by T. A. Sebeok. (The Hague: Mouton), vol. 12.

Lisker, L. and A. S. Abramson. (1971) Distinctive features and laryngeal control. Language 47, 767-785.

MacNeilage, P. F. (1970) Motor control of serial ordering of speech. Psychol. Rev. 77, 182-196.

MacNeilage, P. F. (1972) Speech physiology. In Speech and Cortical Functioning, ed. by J. H. Gilbert. (New York: Academic Press).

Mermelstein, P. (1974) A phonetic-context controlled strategy for segmentation and phonetic labeling of speech. Haskins Laboratories Status Report on Speech Research SR-37/38, 191-197.

Millar, J. B. and W. A. Ainsworth. (1972) Identification of synthetic isolated vowels and vowels in h-d context. Acoustica 27, 278-282.

Miller, G. A., G. A. Heise, and W. Lichten. (1951) The intelligibility of speech as a function of the context of the test materials. J. Acoust. Soc. Amer. 41, 329-335.

Nooteboom, S. G. (1970) The target theory of speech production. IPO Annual Progress Report (Institute for Perceptual Research, Eindhoven, Holland) 5, 51-55.

Nooteboom, S. G. and I. Slis. (1970) A note on the degree of opening and the duration of vowels in normal and "pipe" speech. IPO Annual Progress Report (Institute for Perceptual Research, Eindhoven, Holland) 5, 55-58.

Peterson, G. E. (1951) The phonetic value of vowels. Language 27, 541-553.

Peterson, G. E. (1961) Parameters of vowel quality. J. Speech Hearing Res. 4, 10-29.

Peterson, G. E. and H. L. Barney. (1952) Control methods used in the study of the vowels. J. Acoust. Soc. Amer. 24, 175-184.

Peterson, G. E. and I. Lehiste. (1960) Duration of syllabic nuclei in English. J. Acoust. Soc. Amer. 32, 693-703.

Peterson, G. E., W. S. T. Wang, and E. Sivertsen. (1958) Segmentation techniques in speech synthesis. J. Acoust. Soc. Amer. 30, 739-742.

Rand, T. C. (1971) Vocal tract size normalization in the perception of stop consonants. Haskins Laboratories Status Report on Speech Research SR-25/26, 141-146.

Schatz, C. (1954) The role of context in the perception of stops. Language 30, 47-56.

Shaw, R., M. McIntyre, and W. Mace. (1974) The role of symmetry in event perception. In Perception: Essays in Honor of J. J. Gibson, ed. by R. B. MacLeod and H. L. Pick. (Ithaca, N. Y.: Cornell University Press).

Shaw, R., and J. Pittenger. (in press) Perceiving the face of change in changing faces. In Perceiving, Acting, and Comprehending: Toward an Ecological Psychology, ed. by R. Shaw and J. Bransford. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Stevens, K. N. (1972) The quantal nature of speech: Evidence from articulatory-acoustic data. In Human Communication: A Unified View, ed. by E. E. David, Jr., and P. B. Denes. (New York: McGraw-Hill).

Stevens, K. N. and A. S. House. (1963) Perturbation of vowel articulations by consonantal context: An acoustical study. J. Speech Hearing Res. 6, 111-128.

Stevens, S. S. and J. Volkmann. (1940) The relation of pitch to frequency: A revised scale. Amer. J. Psychol. 53, 329-353.

Strange, W., R. Verbrugge, and D. Shankweiler. (1974) Consonant environment specifies vowel identity. Haskins Laboratories Status Report on Speech Research SR-37/38, 209-216.

Summerfield, A. W. and M. P. Haggard. (1973) Vocal tract normalization as demonstrated by reaction times. Speech Perception, Report on Speech Research in Progress (Psychology Department, The Queen's University of Belfast) Series 2, no. 3, 1-26.

Verbrugge, R., W. Strange, and D. Shankweiler. (1974) What information enables a listener to map a talker's vowel space? Haskins Laboratories Status Report on Speech Research SR-37/38, 199-208.

Watson, J. B. (1924) Behaviorism. (New York: Norton).

Wickelgren, W. A. (1969) Context-sensitive coding, associative memory, and serial order in (speech) behavior. Psychol. Rev. 76, 1-15.

"Coperception": A Preliminary Study

Bruno H. Repp*

ABSTRACT

The present paper defines "coperception," in analogy to coarticulation, as the influence of one segment on the perception of another segment in an utterance. The appropriate measure for this influence is reaction time, in one of several possible tasks (classification, "same-different" judgments, monitoring). The factors that may govern coperception are discussed in terms of three (not mutually exclusive) hypotheses: (1) temporal integration, (2) perception-production correlations, and (3) perceptual units. A preliminary study is reported that demonstrates coperception of a stop consonant with the final vowel in VCV utterances, although the former was partially independent of the latter at the acoustical level. "Same-different" judgments about the consonants in two successive VCVs were influenced by the similarity between the final vowels. This result extends similar findings obtained by others in CV syllables. Some methodological issues in the investigation of coperception are discussed.

INTRODUCTION

Pisoni and Tash (1974) demonstrated that, at the level of "same-different" judgments, the consonant and the vowel in CV (stop-vowel) syllables are not perceptually independent of each other. When the consonants in two successive CV syllables were to be compared, "same" judgments were faster when the vowels were also the same (/ba/-/ba/) than when they were different (/ba/-/bæ/), and "different" judgments were faster when the vowels were different (/ba/-/dæ/) than when they were the same (/ba/-/da/). Similar effects were exerted by the consonants when the identity of the vowels was to be judged. Wood and Day (1975) obtained precisely the same results in a different but related paradigm, speeded classification: both consonant and vowel classifications (binary-choice reaction times) were faster when the irrelevant phoneme was held constant than when it varied randomly.

---

*University of Connecticut Health Center, Farmington.

Acknowledgment: The assistance of Millicent Winston in several stages of the experiment is gratefully acknowledged. Thanks are further due to Haskins Laboratories for the use of their facilities and to the Psychology Department of the University of Connecticut at Storrs for permission to use their equipment and to draw on their subject pool. Chris Darwin provided valuable assistance there. This research was supported by NIH Grant T22 DE00202 to the University of Connecticut Health Center.

[HASKINS LABORATORIES: Status Report on Speech Research SR-42/43 (1975)]

Clearly, neither "same-different" judgments nor classification responses are based exclusively on abstract, independent phonemic codes. These codes, which are subjectively so salient, are presumably higher cognitive constructs that are not directly accessed in rapid perceptual judgments (cf. also Savin and Bever, 1970). The phonemic codes in immediate perception, while permitting error-free judgments, seem to remain "context-sensitive" (Wickelgren, 1969) by the criterion of response latency. The characteristics and limits of this context sensitivity are worth exploring.

The stimuli used by Pisoni and Tash (1974) and by Wood and Day (1975) represented optimal conditions for context sensitivity to emerge. In stop-vowel syllables, the phonemes are not only contiguous but also in part acoustically dependent on each other: the formant transitions transmit information about the consonant and the vowel in parallel. Context sensitivity as a consequence of parallel information transmission at the acoustic level is not a greatly surprising finding. More interesting cases arise when the phonemes in question are independent at the acoustic level or even separated by silence or intervening information. Pisoni and Tash (1974), at least, seem to expect perceptual independence in this case: "If the information were not transmitted in parallel form we would not expect differences in consonants to affect the vowel decision and differences in vowels to affect the consonant decision" (p. 134). However, this conception may be too narrow. Before I proceed to some empirical data, let me propose a new term.

Because of its analogy to the articulatory phenomenon of coarticulation, the perceptual phenomenon under investigation will be termed "coperception." Coperception exists whenever the perception (in terms of the speed or the accuracy of certain judgments) of a particular portion or segment of the speech signal is influenced by a preceding or following portion or segment. Coarticulation, of course, is defined as the influence of one segment on the articulation of another segment (Daniloff and Hammarberg, 1973). Coarticulation may extend over several (acoustic or phonetic) segments, and the same may well be true of coperception. Both phenomena have limits in terms of a maximal time interval or number of segments over which they can extend. Both may work in a forward (left-to-right) and/or in a backward (right-to-left) direction. While left-to-right coarticulation may involve carryover and inertia within the articulatory system, in addition to centrally planned components, right-to-left coarticulation reflects only articulatory planning and anticipation (Daniloff and Hammarberg, 1973) and therefore is perhaps the more interesting effect of the two. Similarly, right-to-left coperception is probably more interesting (or less confounded) than left-to-right coperception. In the latter, the contextual information has already entered the system and may affect subsequent judgments simply because of its presence in some short-term store, or because it is still being processed, or because it has "preset" some relevant response mechanism or criterion--in short, because of its "perceptual inertia." In right-to-left coperception, on the other hand, the biasing context follows the segment to be judged, so that contextual effects will reflect only the size of the perceptual segment, defined as a particular time span or as a particular number or constellation of segments, over which the speech perception mechanism integrates before a decision ("same-different" or classification) can be reached.

As can be seen, the formal analogy between coarticulation and coperception is nearly complete. Whether there is any functional analogy, in terms of the constraints involved or in terms of more specific relationships between adjacent sounds, remains to be investigated. The most obvious question to ask is whether coperception can be found in the absence of coarticulation in the same signal, or vice versa. This may be investigated by examining the coperception of segments that do not show coarticulation, or by looking for evidence of coarticulation between segments whose perceptual independence has been established. The present study had a more modest goal: by using synthetic speech, coarticulation effects were partially removed from utterances that tend to show such effects in natural speech. The question was whether perception would take advantage of this opportunity and show independence, or whether coperception would nevertheless be found, which could then be explained either by reference to articulatory constraints or to more general temporal limits of perception.

More specifically, the present study used the "same-different" paradigm with stimuli that went just one step beyond the CV syllables employed by Pisoni and Tash (1974). Vowel-consonant-vowel utterances were synthesized *such that the acoustic segment preceding the stop closure was invariant with the final vowel*. Since this acoustic segment, which is perceived in isolation as a VC syllable, is sufficient for recognition of the stop consonant, the speech perception mechanism was offered an opportunity to show independence of the final vowel in judgments of the consonant. On the other hand, Öhman (1966) has shown that, in natural speech, the implosive transitions (which precede the closure) do vary with the final vowel to some degree. The explosive transitions (which follow the closure) were dependent on the final vowel, of course, and corresponded to those of a CV syllable. If the perceptual mechanism integrates implosive and explosive transitions across the closure (50 msec of silence), coperception should occur. If, on the other hand, a response is initiated as soon as there is minimal information present to reach a decision, perceptual independence may be found. As an additional aspect, the present study used not two but four different final vowels, in order to investigate the effects that their mere *similarity* may have on "same-different" judgments.

METHOD

Subjects

Twelve female University of Connecticut undergraduates received course credit for their participation.

Stimuli

Eight VCV utterances, /aba/, /abæ/, /abɛ/, /abi/, and /ada/, /adæ/, /adɛ/, /adi/, were synthesized by rule on a Glace-Holmes synthesizer at the University of Connecticut (Storrs). All stimuli began with 90 msec of steady-state /a/, followed by 50-msec transitions appropriate for a final /-b/ or /-d/, which did *not* depend on the final vowel. These implosive transitions were followed by 50 msec of silence, then 50 msec of explosive transitions (dependent on the final vowel), and finally one of four steady-state vowels of 100-msec duration. The total VCV duration was 340 msec. The final vowel /-ɛ/ was mistakenly synthesized to be only of 40-msec duration, so that these two utterances lasted only 280 msec. (The results showed little effect of this factor--see below.)

The experimental tape was recorded with the help of the pulse-code-modulation (PCM) system at Haskins Laboratories. The tape contained first a random list of 80 single VCVs, 10 replications of each individual stimulus. Each VCV was preceded by a 100-msec warning tone (a synthetic nonspeech signal), which came on 500 msec prior to VCV onset. The interstimulus interval (between VCV offset and the onset of the next warning tone) was 4 sec. This practice series was followed by five blocks of 64 randomized test pairs (i.e., five replications of all possible pairings of the eight stimuli). The onset-onset interval between the two VCVs in a pair was 490 msec. The first VCV was again preceded by a warning tone, and the interpair interval was 4 sec.

PROCEDURE

The subjects were tested individually in two sessions on different days lasting approximately 1 1/2 hours altogether. The subject listened over Grason-Stadler (Telephonics) TDH-39 earphones and operated a toggle switch with her preferred hand. After the toggle switch and the nature of the stimuli were explained, the subject listened to the practice series, with the instructions to classify the consonant in each VCV as rapidly and as accurately as possible by moving the toggle switch in the appropriate direction. The two positions of the switch (toward and away from the body) were labeled "B" and "D"; the assignment of these labels to the positions was counterbalanced over subjects. Subsequently, the five blocks of pairs were presented (with rest pauses between blocks, as required). The labels were changed to "SAME" and "DIFF" (likewise counterbalanced over subjects), and the subject was instructed to judge whether the consonants in the two VCVs in a pair were the same or different as rapidly and as accurately as possible. The variation in the final vowel was to be ignored. In the second session, the subject repeated the five experimental blocks.

Because of equipment malfunction, half of the subjects listened to the stimuli monaurally; some blocks were presented to the left ear and some to the right ear. The remaining subjects listened binaurally, as intended. The tapes were played back on a high-quality tape recorder (Crown 800) and passed through a Lafayette La-375 solid-state amplifier. The intensity was set at a comfortable listening level. Reaction times were recorded by a Hunter MFG 1521 digital millisecond timer. The timer was triggered by a Grason-Stadler E7300A-1 voice-operated relay key, which in turn was activated by the onset of the warning tone on the tape. The timer was manually reset by the experimenter after she recorded the reaction time.

Although reaction times were originally measured from the onset of the warning tone, a constant was subsequently subtracted, such that the reaction times were measured from the offset of the implosive transitions (the onset of the silent closure period) of the second VCV in a pair (of the single VCV in the classification task). Median reaction times were calculated for the ten replications of each practice stimulus, and for the five replications of each individual experimental pair in each session, omitting errors. These medians formed the basic data for further analysis, which was in terms of averages.

RESULTS

Classification

Table 1 shows the average median reaction times for classifying the medial consonants as "B" or "D." It can be seen that the reaction times for "D" were slightly longer than those for "B," but there was a really substantial difference only in one context, /-i/: /abi/ had the fastest and /adi/ the slowest latencies of all stimuli. This specific interaction reached significance (*p*<.03).

TABLE 1: Average median classification reaction times (msec) and error percentages (in parentheses) for /ab-/ ("B") and /ad-/ ("D") in four different contexts.

	/-a/	/-æ/	/-ɛ/	/-i/	Mean
/ab-/	331 (4.2)	334 (3.4)	336 (0.8)	322 (6.7)	331 (3.8)
/ad-/	334 (1.7)	337 (1.7)	346 (5.8)	377 (3.4)	349 (3.2)

The error rates are also shown in Table 1. No stimulus caused particular difficulties and all VCVs were highly identifiable, especially considering that they were synthetic. The differences in error rates between individual stimuli were almost certainly random.

"Same-Different" Judgments

To simplify the statistical and graphical analysis of the data, a "similarity" dimension was defined on the final vowels. Pairs of vowels that are neighbors in the classical vowel triangle (or quadrilateral) were considered "similar," while other pairs were considered "dissimilar." The similar pairs were: /-a/+/-æ/, /-æ/+/-ɛ/, and /-ɛ/+/-i/; the dissimilar pairs were /-a/+/-ɛ/, /-a/+/-i/, and /-æ/+/-i/. The median reaction times were subjected to a four-way analysis of variance (Bock's, 1975, pseudo-multivariate method for repeated measurements), with the factors: (a) Type of Response ("same" versus "different"), (b) Vowel Context (identical - similar - dissimilar), (c) Consonant of First VCV (B versus D), and (d) Sessions (first versus second). The reason for the third factor will become evident below. In addition, three-way analyses of variance were conducted on "same" and "different" latencies separately.

The results are shown in Figure 1. "Same" latencies were significantly faster than "different" latencies (*p*<.002), and there was also a significant main effect of Vowel Context (*p*<.03). Both results make little sense, however, in view of the striking interaction between the two factors (*p*<.0004): "same" latencies were much faster than "different" latencies when the final vowels were identical, somewhat faster when the vowels were similar, and a bit slower when the vowels were dissimilar. (The number of individual subjects showing faster "same" latencies--by at least 10 msec--in the three contexts was 12, 9, and 2, respectively.) The two component contrasts of the interaction, which contrasted identical with nonidentical vowels and similar with dissimilar vowels, respectively, were both significant (*p*<.0001 and *p*<.002, respectively).

The vowel context had a striking effect on "same" latencies (*p*<.002), while its effect on "different" latencies was less pronounced (*p*<.05) and in the opposite direction. "Same" latencies were significantly faster in identical contexts than in nonidentical contexts (*p*<.0004, shown by all 12 subjects), while "different" latencies were significantly slowed down in identical contexts

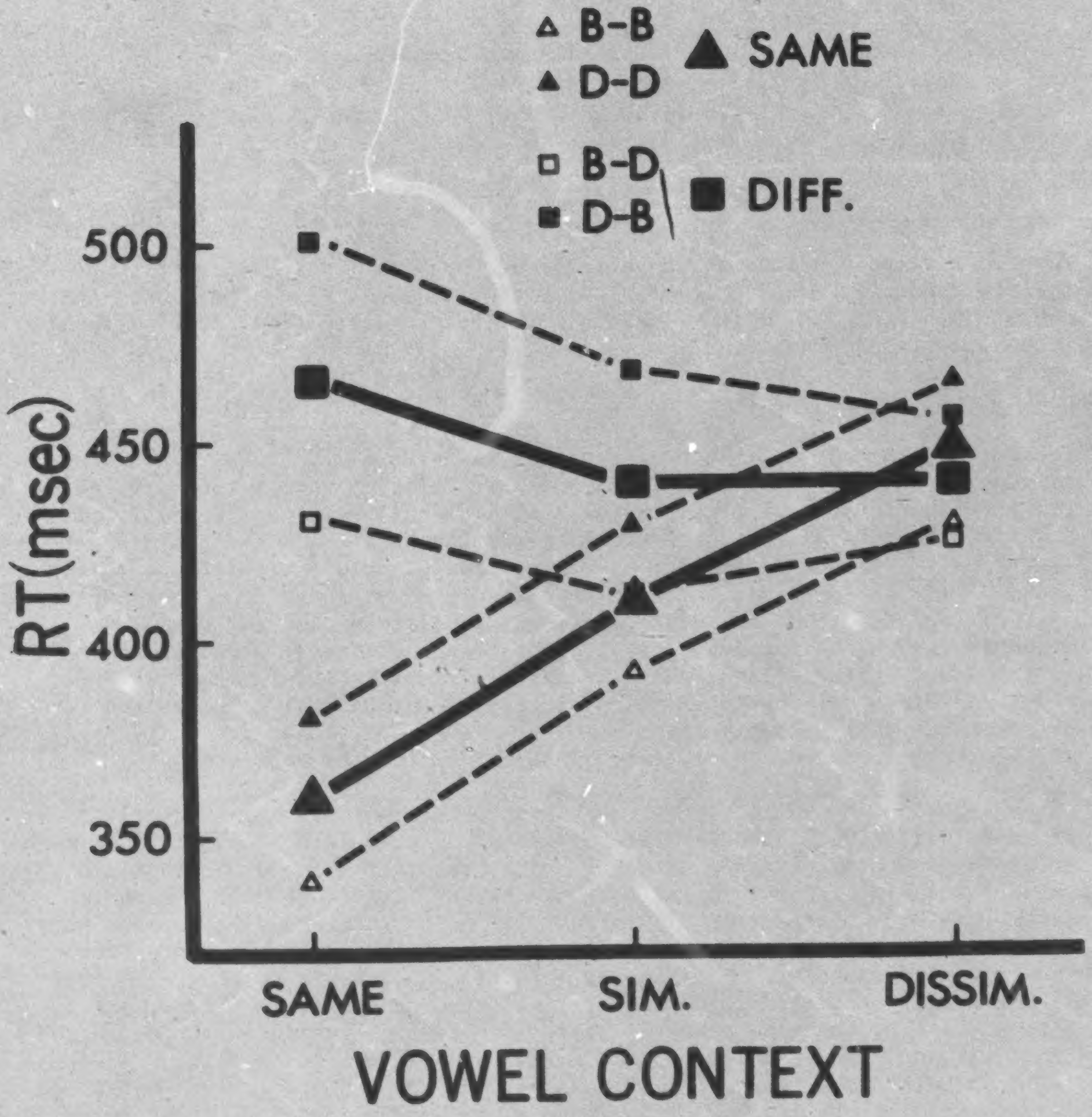

Figure 1: Average median reaction times (RT) of "same" and "different" judgments in three different classes of vowel context. The small symbols represent particular sequences of the consonants being judged.

(p<.02, shown by 8 subjects). Vowel similarity, on the other hand, affected only "same" responses, which were much faster in similar contexts (p<.0009, shown by 10 subjects); "different" responses showed no difference at all. However, four subjects surprisingly showed *faster* "different" latencies in similar contexts than in dissimilar contexts, while three subjects showed a difference in the opposite direction.

Figure 1 shows further that the nature of the consonant in the *first* VCV exerted a large effect on the response times: they were much slower when the first consonant was a D than when it was a B. Because of high variability, the significance level of the effect was not very high (p<.02). The difference was only marginally significant for "same" latencies alone (p<.06), although the difference was in the same direction for all subjects, but more consistent for "different" latencies (p<.007), owing to enormously large differences for some subjects (but two subjects showed an inverted effect). The nature of the second consonant apparently had no influence on the reaction times (cf. Table 2).

Finally, there was also a significant decrease of reaction times with practice, i.e., from the first to the second session (p<.002). However, there were no substantial changes in the pattern of results. The effect of identical contexts on the speed of "same" responses was somewhat less pronounced in the second session (p<.04) but still very striking. The same was true with respect to the effect of the first consonant (p<.05).

The classification of the vowels into similar and dissimilar pairs may have been somewhat arbitrary and was partially confounded with the individual vowels. Also, one of the vowels, /-ɛ/, had a shorter duration than the others. Therefore, the detailed results are shown in Table 2.

TABLE 2: Average median latencies (msec) for "same" (upper-left and lower-right quadrants) and "different" (upper-right and lower-left quadrants) judgments. Latencies for identical vowel contexts are underlined.

First VCV		Second VCV								
		/ab-/				/ad-/				
		/-a/	/-æ/	/-ɛ/	/-i/	/-a/	/-æ/	/-ɛ/	/-i/	Mean
/ab-/	/-a/	332	408	404	451	399	383	381	408	396
	/-æ/	385	393	369	447	389	416	411	430	405
	/-ɛ/	418	383	301	391	454	432	472	428	410
	/-i/	405	454	414	327	439	440	425	434	417
/ad-/	/-a/	509	469	464	472	399	424	453	462	457
	/-æ/	466	528	460	467	409	408	410	457	451
	/-ɛ/	464	494	486	453	456	420	354	443	446
	/-i/	439	432	463	474	437	524	465	356	449
	Mean	427	445	420	435	423	431	421	427	429

There are only a few atypical results at the level of individual stimulus pairs. For example, in /-æ/ context, "same" latencies were *not* faster when the vowels were identical; there is no obvious explanation for this result. The fastest "same" latencies and, in one quadrant, the slowest "different" latencies occurred in identical /-ε/ contexts. This is interesting, since it may reflect the additional effect of final vowel duration, a nonphonetic variable. There were no effects reflecting the different classification times observed in /-i/ context during the practice series.

The pattern of *error* percentages, when plotted as in Figure 1, was so similar to that of the latencies that a separate figure is superfluous. In other words, the error frequencies were highly correlated with the latencies of correct responses and showed the same effects of vowel context, as well as the influence of the first consonant. While it is well-known that the errors and the latencies in "same-different" judgments are positively correlated, the degree of similarity found here was surprising. The error percentages of the individual subjects varied widely, from 0.8 to 17.2 percent, but eight of the subjects remained below 5.0 percent. Nevertheless, all subjects showed the same basic pattern of reaction times, so that latencies and errors merely appear to be alternative reflections of the same underlying processes.

DISCUSSION

The basic outcome of the present study is clear: in the context of a "same-different" judgment task, the perception of the consonant in VCV utterances is not independent of the final vowel, although there is a part of the acoustic signal that is not dependent on the final vowel and sufficient for recognition of the consonant. This is a clear case of right-to-left "coperception," as defined in the Introduction, and somewhat less trivial than coperception in CV syllables, where the acoustic cues for the stop consonant are largely dependent on the following vowel.

There are several possible explanations for why coperception occurred in the present situation; they are all interesting and not at all mutually exclusive: (1) *Temporal integration*: One possibility is that the information that was coperceived with the consonant occurred within the time span over which the perceptual mechanism integrates, so that it could not possibly be ignored. This is plausible, since vowel-independent and vowel-dependent transitions were separated by merely 50 msec of silence. On the other hand, the integration or temporal storage limits of the speech processor seem to be on the order of 250 msec, according to estimates from backward-masking paradigms (Massaro, 1972, 1974; Repp, 1975a); or perhaps only about half as long, according to estimates from certain other dichotic masking studies (e.g., Pisoni and McNabb, 1974), from dichotic discrimination errors (Repp, 1975b), or from the perception of dichotic pulse trains (Huggins, 1974). Both estimates are sufficient to explain the present results. The next step would be to try to exceed the limits of the temporal integration period by increasing the duration of the silent closure period. If it is sufficiently long (perhaps about 150 msec), geminate consonants will be heard, i.e., /ab-ba/ instead of /aba/ (Dorman, Raphael, Liberman, and Repp, 1975). It is a reasonable guess that coperception will disappear precisely at that point, but this remains to be tested. If the hypothesis is confirmed, it would provide support for Massaro's (1972) conjecture of a 250-msec integration period in speech perception, since the duration of the transitions that are integrated will have to be added to the silent interval to obtain the maximal integration interval.

However, there may be other factors at work than purely temporal ones. (2) Perception-production correlations: One relevant finding is that, in natural speech, the implosive transitions in VCVs show some degree of coarticulation with the final vowel (Öhman, 1966). Since speech perception has often been shown to take into account the dynamics and constraints of speech production, we may have found another manifestation of this functional symbiosis. In other words, coperception may occur whenever coarticulation normally occurs, even if there is no actual evidence of coarticulation in the speech signal (because the synthetic speech has been deliberately generated without it, as in the present study). It will be difficult to produce unequivocal support for this hypothesis, but it may be critically tested, for example, by examining whether coperception occurs across synthetic consonant clusters followed by vowels that include /u/ which, in normal speech, produces anticipatory lip-rounding throughout the whole cluster (Kozhevnikov and Chistovich, 1965; Daniloff and Moll, 1968).

(3) "Perceptual units": There is a third possible explanation for coperception in VCVs, and it concerns perceptual preferences for certain constellations of phonetic segments. Kozhevnikov and Chistovich (1965) have postulated that the CV syllable forms the basic perceptual (and articulatory) unit, so that a VCV utterance should be perceptually parsed into V-CV. Of course, this is consistent with the present results. There are obvious ways in which this perceptual unit hypothesis could be further tested. For example, it predicts that there should be less or no coperception between the first vowel and the consonant in VCVs when judgments about the consonant are to be made. This is an interesting hypothesis to test, since the vowel-dependent transitions precede the vowel-independent transitions in this case, so that the predicted outcome would be somewhat counterintuitive. Another way of testing the influence of segmental factors while holding temporal relationships constant would be to replace the explosive transitions in a VCV with transitions appropriate for a different consonant, e.g., /ab-da/. However, it is known that, in this case, the perception of the first consonant would be impaired (Dorman et al., 1975), a result that in itself is consistent with the result of coperception in the present VCVs and with the perceptual unit hypothesis. In order to preserve the first consonant, the closure interval would have to be extended to about 100 msec; but the question of whether coperception occurs may then still be asked, also with regard to perceptual interactions between the two consonants.

Coperceptual phenomena may be influenced by further variables not yet mentioned. Suprasegmental factors, such as the relative stress, duration, or intonation of the two vowels in a VCV utterance, may play a role. Syllable, morpheme, or even syntactic boundaries may affect coperception. The kind of instructions and the task may be important, too. For example, in speeded classification (Wood and Day, 1975), the listeners may be comparing the stimuli with more abstract codes in their brain than in the "same-different" judgment task where the first stimulus in a pair must be encoded before it is compared with the second stimulus. If the second stimulus occurs before the encoding of the first is complete--and the process of phonemic abstraction may take more time than is usually allowed--coperception in some form is likely to result. The interval between the stimuli to be compared may therefore be an important variable. Other important issues, such as that of the relative discriminability of parts of the signal, have been discussed by Wood and Day (1975). Another related paradigm that may be employed to investigate coperception is monitoring--rapid detection of phonemic targets in a series of signals that either contain the target or do not. This paradigm has been primarily connected with discussions of the perceptual unit hypothesis in the past, since it can also be easily applied to larger-size units such as syllables or words (Savin and Bever, 1970; Lehiste, 1972; McNeill and Lindig, 1973). Each of the three paradigms discussed here may be applied to investigate coperception, but the specific demands of each task will have to be studied carefully.

Besides demonstrating the basic phenomenon of coperception, the present experiment yielded an interesting result that must be taken into account in any model of "same-different" judgments: the mere similarity of the final vowels influenced the speed of "same" judgments. (The influence on "different" judgments was not consistent.) It is known that the speed of "different" judgments increases with increasing dissimilarity of the stimuli being compared, and, similarly, the speed of "same" judgments (including incorrect responses) decreases with increasing dissimilarity of the stimuli (e.g., Repp, 1975b). Given that a consonant and a vowel are coperceived, the similarity between the vowels affects "same" judgments about the consonant. This rules out any discrete model that admits only decisions about identity and nonidentity (Pisoni and Tash, 1974). Rather, the response latencies reflect an underlying perceptual continuum, or at least a more fine-grained discrete representation such as feature matrices. Why the "different" latencies were not affected by vowel similarity is not obvious, and since different subjects showed differences in opposite directions, no interpretation will be attempted here.

A truly puzzling finding is the influence of the first consonant in a pair on both "same" and "different" reaction times. (At first, I suspected this effect to be due to an error in data transcription, but this seems to have been ruled out.) A difference due to the second consonant would have been infinitely more plausible, since the decision time includes the time to encode the second stimulus. Of course, the situation is ambiguous: it may equally well represent an interaction of the second consonant with the kind of judgment made, but this does by no means facilitate its interpretation. Moreover, there were no differences of a similar extent in the classification latencies for single syllables. This effect should definitely be replicated, since it still may be due to some undiscovered artifact.

It has often been suggested that the syllable is the basic perceptual unit (e.g., Savin and Bever, 1970; Massaro, 1972). However, since the syllable is usually not precisely defined, this statement is somewhat circular. Linguistic definitions of the syllable may not be directly relevant to perception, and specific proposals, such as the primacy of the CV syllable (Kozhevnikov and Chistovich, 1965) need to be further tested. Research on coperception provides a discovery procedure for the units in speech perception--for example, coperception (especially from right to left) should not occur across "syllable boundaries." Systematic research on the limits of coperception promises to provide important information about the processes involved in speech perception, and perhaps about the perception of time-varying signals in general.

REFERENCES

Bock, R. D. (1975) *Multivariate Statistical Methods in Behavioral Research.* (New York: McGraw-Hill).

Daniloff, R. G. and R. E. Hammarberg. (1973) On defining coarticulation. J. Phonetics 1, 239-248.

Daniloff, R. G. and K. Moll. (1968) Coarticulation of lip-rounding. J. Speech Hearing Res. 11, 707-721.

Dorman, M. F., L. J. Raphael, A. M. Liberman, and B. H. Repp. (1975) Masking-like phenomena in speech perception. J. Acoust. Soc. Amer., Suppl. 57, 48(A). [Also in Haskins Laboratories Status Report on Speech Research SR-42/43 (this issue).]

Huggins, A. W. F. (1974) On perceptual integration of dichotically alternated pulse trains. J. Acoust. Soc. Amer. 56, 939-943.

Kozhevnikov, V. A. and L. A. Chistovich. (1965) Speech, Articulation, and Perception. (Washington, D.C.: Joint Publications Research Service).

Lehiste, I. (1972) The units of speech perception. In Speech and Cortical Functioning, ed. by J. H. Gilbert (New York: Academic Press).

Massaro, D. W. (1972) Preperceptual images, processing time, and perceptual units in auditory perception. Psychol. Rev. 79, 124-145.

Massaro, D. W. (1974) Perceptual units in speech recognition. J. Exp. Psychol. 102, 199-208.

McNeill, D. and K. Lindig. (1973) The perceptual reality of phonemes, syllables, words, and sentences. J. Verbal Learn. Verbal Behav. 12, 419-430.

Öhman, S. E. G. (1966) Coarticulation in VCV utterances: Spectrographic measurements. J. Acoust. Soc. Amer. 39, 151-168.

Pisoni, D. B. and S. D. McNabb. (1974) Dichotic interactions of speech sounds and phonetic feature processing. Brain Lang. 1, 351-362.

Pisoni, D. B. and J. Tash. (1974) "Same-different" reaction times to consonants, vowels, and syllables. In Research on Speech Perception, Progress Report No. 1. (Department of Psychology, Indiana University).

Repp, B. H. (1975a) Dichotic forward and backward "masking" between CV syllables. J. Acoust. Soc. Amer. 57, 483-496.

Repp, B. H. (1975b) Distinctive features, dichotic competition, and the encoding of stop consonants. Percept. Psychophys. 17, 231-240.

Savin, H. B. and T. G. Bever. (1970) The nonperceptual reality of the phoneme. J. Verbal Learn. Verbal Behav. 9, 295-302.

Wickelgren, W. A. (1969) Context-sensitive coding, associative memory, and serial order in (speech) behavior. Psychol. Rev. 76, 1-15.

Wood, C. C. and R. S. Day. (1975) Failure of selective attention to phonetic segments in consonant-vowel syllables. Percept. Psychophys. 17, 346-350.

Dichotic "Masking" of Voice Onset Time*

Bruno H. Repp+

ABSTRACT

When consonant-vowel (CV) "targets" are presented to one ear and isolated vowel "masks" to the other ear, the perception of the voicing feature of the target is biased by the temporal relationship between target and mask which acts like a "pseudo voice onset time" and competes with the actual voice onset time of the target. The effect is especially strong when there is high *a priori* uncertainty about the voicing category of the target, and it is more pronounced when the masking vowel lags behind than when it leads in time. The bias is stronger when the masking vowel is the same as the vowel of the target, but it is present with a different vowel mask as well. There tends to be a right-ear advantage that is strongest when the masking vowel leads in time. The fundamental frequency at masking vowel onset has an additional influence: the lower it is, the stronger is the tendency to identify the target as voiced. Two simple additive models of the vowel "masking" effect are rejected. The complexity of the effect suggests dichotic interaction at the phonetic level. The present experiments demonstrate that a temporal relationship may be masked by another temporal relationship, and that the voice onset time across ears may act as a phonetic cue.

INTRODUCTION

In a study of the dichotic masking of consonants by vowels (Repp, 1975b) three main results were reported, all of which were more or less unexpected and were revealed only by a rather detailed breakdown of the data. The task consisted of identifying consonant-vowel (CV) syllables (from the set: /ba/, /da/,

---

*To be published in The Journal of the Acoustical Society of America.

+University of Connecticut Health Center, Farmington, Conn.

Acknowledgment: This research would not have been possible without the generous hospitality and support of Haskins Laboratories where the experiments were prepared and conducted. I thank especially Alvin Liberman for his invitation to Haskins and for his continuous encouragement, and Tom Gay for his support and for comments on the paper. This research was supported by NIH Grant No. T22 DE00202 to the University of Connecticut Health Center.

/ga/, /pa/, /ta/, /ka/) presented to the right ear while isolated vowels (/a/) were simultaneously presented to the left ear at varying stimulus onset asynchronies (SOAs) with respect to the CV syllables. The three results were: (1) The vowel "masks" interfered primarily with the perception of voicing in the opposite ear and only little with the perception of place of articulation. (2) When the onset of the masking vowel preceded the onset of the target syllable, there was an increased tendency to give voiced responses (/ba/, /da/, /ga/), but when the vowel lagged behind, there was a rapid shift in the opposite direction and voiceless responses dominated (/pa/, /ta/, /ka/). (3) Of two vowel masks with different pitch contours, the vowel with the higher fundamental frequency led to more voiceless responses than the one with the lower pitch, and this effect seemed to be independent of the bias due to the temporal relationship between target and mask.

One possible explanation proposed for these findings was that the auditory information from the two ears is combined at a relatively early perceptual stage and that the interfering vowel masks the voice onset time (VOT) of the target syllable by substituting another voicing cue: the perceptual mechanism mistakenly accepts the SOA between target syllable and masking vowel as the VOT of the target syllable. (Naturally, the onset of the vowel mask implies the onset of voicing, i.e., of fundamental frequency.)

The present two experiments attempt to replicate and extend the earlier fortuitous findings (Repp, 1975b) in a somewhat different, methodologically improved paradigm. One restrictive feature of the previous study was that the target syllables were presented at relatively low intensities (below asymptotic intelligibility) and that the observed effects of the masking vowel were pronounced only when it was at a higher intensity than the target syllables. In the two studies to be described here, all stimuli were presented at higher, and equal, intensities, but the VOT of the target syllables was systematically varied. By presenting target syllables from a VOT continuum (Lisker and Abramson, 1967), the actual VOT of the target syllable was juxtaposed to the "pseudo-VOT" presumably simulated by the SOA between target and mask. It was expected that syllables with a VOT close to the category boundary, that is, with high uncertainty about their phonetic category, would be affected most.

The effect of fundamental frequency was reinvestigated in the first experiment by using three different pitch contours for the masking vowels. In addition to assessing the effects of SOA and pitch in a more precise fashion, the present experiments examined laterality effects by varying the ear to which the targets were presented (which was fixed in Repp, 1975b). The vowel-masking paradigm is interesting with respect to the ear advantages often found in dichotic experiments: it lies somewhere between the familiar paradigm of competing CV syllables for which a right-ear advantage is usually obtained (for example, Studdert-Kennedy and Shankweiler, 1970) and that of competing vowels, which typically show no ear difference (e.g., Darwin, 1971). More specifically, we seem to be dealing here with the perception of a single feature, voicing, which is represented at the acoustic level by a temporal relationship, VOT. Likewise, the competing information, the "pseudo-VOT" due to SOA, is a temporal relationship (between the two ears, in this case). This kind of perceptual competition is a novel situation, but a right-ear advantage may be expected, since there is some evidence that the left hemisphere is superior for fine temporal discriminations (e.g., Efron, 1963; Halperin, Nachshon, and Carmon, 1973), apart from its specifically linguistic capacities.

EXPERIMENT I

Method

Subjects. Of the ten subjects who participated, six were paid volunteers recruited through advertisements on Yale University campus, and four were unpaid volunteers from the staff of Haskins Laboratories. One of the latter was an experienced listener; his results were included, since they did not differ systematically from those of the other nine subjects who had rarely or never listened to synthetic speech before. There were seven females (hence, a subject will be referred to as "she" in the following) and three males. One female was left-handed. All subjects were native speakers of English and claimed to have normal hearing.

The results of these ten subjects will be compared with those for a highly experienced listener (BHR, the author), who repeated the experiment six times, in order to assess the effects of experience and practice in this particular task, and to produce stable data for a single listener. The author is right-handed and a native speaker of the Austrian dialect of German. (While the results for BHR will be reported below, his data are not shown or included in any of the figures.)

Stimuli. The target stimuli were ten synthetic syllables taken from two VOT continua, /bɛ/-/pɛ/ and /dɛ/-/tɛ/. (For convenience, they will be referred to as B-P and D-T continua.) The VOTs of the five syllables from each continuum were 10, 20, 30, 40, and 50 msec, simulated by noise excitation and first-formant cutback prior to voicing onset. All syllables were produced on the Haskins Laboratories parallel formant synthesizer. Their duration was 300 msec, and their fundamental frequency fell linearly from 130 to 90 Hz.

The masking vowels were three steady-state /ɛ/ vowels of 240-msec duration, differing only in their pitch contours. The starting frequencies were 112 Hz (low), 122 Hz (medium), and 136 Hz (high); they all fell linearly to 90 Hz at the end of the vowel.

The dichotic tapes were constructed with the aid of the pulse-code-modulation (PCM) system at Haskins. Each target syllable occurred in combination with each of the three masking vowels at each of five SOAs: -10, 10, 30, 50, and 70 msec. (A negative SOA indicates that masking vowel onset preceded target syllable onset.) Moreover, the target stimuli from the middle of the two continua (VOT = 30)—which were expected to fall closest to the category boundaries—were replicated four times, the adjacent stimuli (VOTs = 20 and 40) twice, and the endpoint stimuli (VOTs = 10 and 50) only once. Thus, the total stimulus set consisted of 2 (target continua) × 10 (5 VOTs: 1 + 2 + 4 + 2 + 1 = 10 replications) × 3 (masking vowels) × 5 (SOAs) = 300 stimuli, which were completely randomized. The interstimulus interval was 3 sec.

Procedure. The subjects were tested individually or in small groups in a single session lasting about 90 minutes. The experimenter (BHR) usually joined the subjects in listening. The stimulus tapes were played back from an Ampex AG-500 tape recorder through a listening station interface to Grason-Stadler (Telephonics) TDH-39 earphones. Output intensity was calibrated on a Hewlett-Packard voltmeter. The average intensities (peak deflections) of the target syllables and the masking vowels were equalized at approximately 73 dB SPL. Usually, the left and right channels were interchanged by means of an electronic switch, but in some cases (because of problems with the switch) by reversing the earphones.

Each subject first listened monaurally to a random series of 150 syllables and tried to identify each as "B", "P", "D", or "T" by writing down one response for each syllable. Subsequently, the whole series of 300 dichotic stimuli was presented, with the targets in the ear that had previously received the monaural stimuli. The subjects were instructed to try to ignore the vowels and identify the syllables as accurately as possible. After a break, the channels were reversed, and a second dichotic series was presented, in a different random order and with the targets in the opposite ear. Finally, another list of 150 monaural syllables was presented in the opposite ear. The sequence of (target) ears was counterbalanced between subjects (and across sessions for BHR).

The purpose of the experiment was explained to the subjects after the session was completed. During the experiment, the subjects were not informed that VOT, SOA, and vowel pitch varied, and none of these variations became really obvious while listening, as gathered from the subjects' remarks and from the author's own impressions. When BHR listened, his attention was not drawn to any of these variables, and he never attempted to "listen for them" or compensate for any of their known effects. Rather, he tended to perform the task passively and automatically, occupying his mind with other things while his hand wrote down the responses, in contrast to the naive subjects who concentrated hard.

Results

Monaural identification. The monaural identification functions for the ten subjects are shown in Figure 1. It is evident that fewer voiced responses were given to the stimuli from the B-P continuum than to the stimuli from the D-T continuum, or, in other words, the dental boundary was farther to the right (by about 6 msec) than the labial boundary, as expected (Lisker and Abramson, 1967). However, two of the ten subjects showed no pronounced difference in their two boundaries. Several subjects had difficulties in hearing "B"s, so that the average B-P identification function did not reach asymptote at the shortest VOT (10 msec).

Not surprisingly, BHR's identification functions were steeper than those of the naive subjects, but they showed the same separation of the two continua. His average category boundaries (the boundaries in individual sessions varied over a range of 5 msec) were found to lie 3 msec farther to the right than the average boundaries of the ten subjects, a highly significant difference in terms of response distributions. This may be a reflection of his native language: German voiceless stops tend to be more aspirated than their English counterparts, and this may lead to a corresponding difference in the perceptual criteria for voicelessness.

Not shown in Figure 1 are the place errors, that is, the confusions between the two continua. For BHR and several of the subjects, such confusions were extremely rare. The great majority of place confusions was contributed by two subjects who frequently substituted dental for labial consonants (one of them only in the first identification series). Parenthetically, it is interesting to note

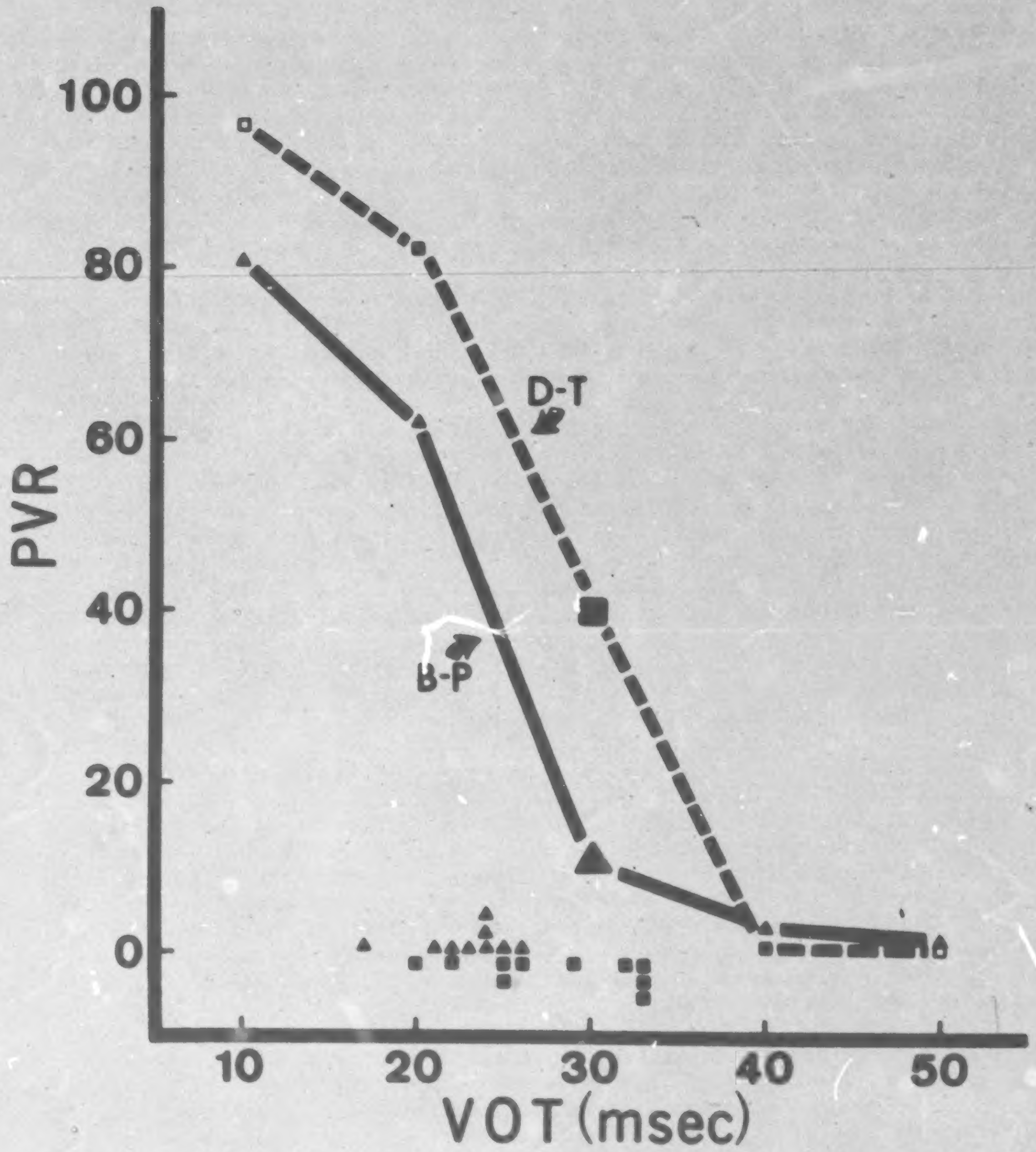

Figure 1: Average percentages of voiced responses (PVR) to monaural stimuli from labial and dental VOT continua in the context /-ε/. The prominence of the individual data points (open - filled - large filled) is in proportion to the number of observations. The small triangles and squares above and below the PVR = 0 level, respectively, represent the distributions of the individual category boundaries (PVR = 50) on the two continua. (The B-P boundary of one subject fell beyond the VOT range shown and is not represented.)

that one of these two listeners (who improved later in the session) showed the labial category boundary in her dental confusion responses. The other subject (who did not improve) seemed to follow the dental boundary in her substitutions.

The effect of SOA. The results of the dichotic condition are shown in Figures 2 and 3 as percentages of voiced responses to the two continua. On each continuum, the five target stimuli are plotted separately, but in the statistical analysis they were treated together in a single weighted percentage score to which an arcsine transformation was applied. Also, because of problems of statistical power, the SOA factor was reduced to two levels by contrasting the two shortest with the two longest SOAs, omitting SOA = 30. The resulting factors in the four-way analysis of variance [Bock's (1975) pseudo-multivariate method for repeated measurements] were: Ears (targets right vs. targets left), Place (B-P vs. D-T), Pitch (of the vowel mask: low-medium-high), and SOA (-10 and 10 vs. 50 and 70). The between-subjects factor, "Order of Target Ears" (left-right vs. right-left) was initially included but omitted after it was found to have had no significant effects. In other words, the effects discussed here did not decrease with practice. Likewise, no change over sessions was observed for BHR.

Figures 2 and 3 show the clear effect that SOA had on the percentages of voiced responses ($p<.0007$). The results of BHR showed the effect to the same extent ($p<.0007$). It was most pronounced for stimuli close to the category boundaries. The overall pattern indicates that the relative sizes of the "positive" and "negative" biases (increase vs. decrease in voiced responses) were dependent on the control score for a stimulus, that is, they were sensitive to the range of variation possible in either direction. In addition, the negative bias seemed to be somewhat more pronounced than the positive bias.

The main effect of Place was significant ($p<.04$): more voiced responses were given to the D-T continuum (cf. Figure 1). This was especially pronounced for BHR ($p<.006$). The Place × SOA interaction was significant for BHR only ($p<.004$): for him, the effect of SOA was stronger on B-P than on D-T, while the ten subjects actually showed an opposite tendency (cf. Figures 2 and 3).

In the figures, it may be noted further that the "positive" effect on the B-P continuum seemed to decrease between SOA = 10 and SOA = -10; this tendency was less pronounced, or absent, on the D-T continuum. For BHR, the "negative" effect also seemed to decrease between SOA = 50 and SOA = 70. In no case, however, did either effect reach the (monaural control) asymptote within the range of SOAs investigated (except for crossing it in the middle range). Stimuli from the ends of the D-T continuum were virtually unaffected by SOA in BHR's data, and Figure 3 shows a related tendency.

The effect of Pitch. The effect of vowel pitch is shown in Figure 4. The curves have been averaged over all ten stimuli, since the Place × Pitch and Place × Pitch × SOA interactions were nonsignificant. As can be seen, the fundamental frequency at vowel onset exerted a strong effect in the expected direction; that is, high pitch led to a relative decrease, and low pitch to a relative increase in voiced responses. The main effect of Pitch was highly significant ($p<.0002$), and so were both its component contrasts, which compared high and low pitch, respectively, with medium pitch. The Pitch effect was equally pronounced for BHR ($p<.002$). In general, high pitch was sufficient to eliminate the positive bias at short SOAs, but low pitch did not eliminate the negative bias at long SOAs.

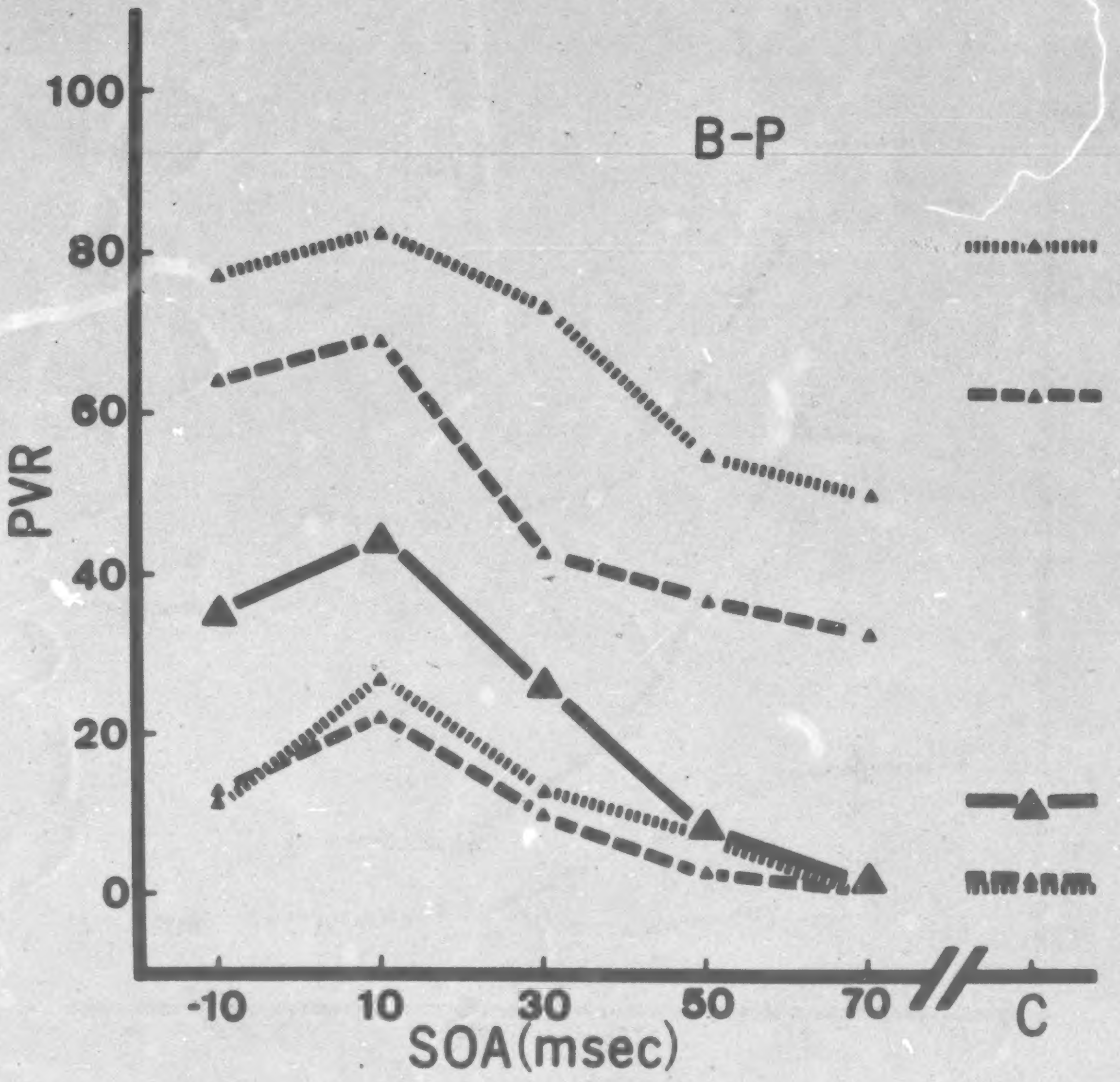

Figure 2: Average percentages of voiced responses (PVR) to the five stimuli from the B-P continuum, in the presence of vowel masks in the other ear at five different SOAs. The prominence of the individual data points and of their connecting lines is in proportion to the number of observations. The monaural control PVRs (from Figure 1) are shown at the right (C). Large symbols represent VOT = 30; small filled symbols, VOT = 20 (top) and 40 (bottom); and small open symbols, VOT = 10 (top) and 50 (bottom). The data are averaged over the Pitch and Ears factors.

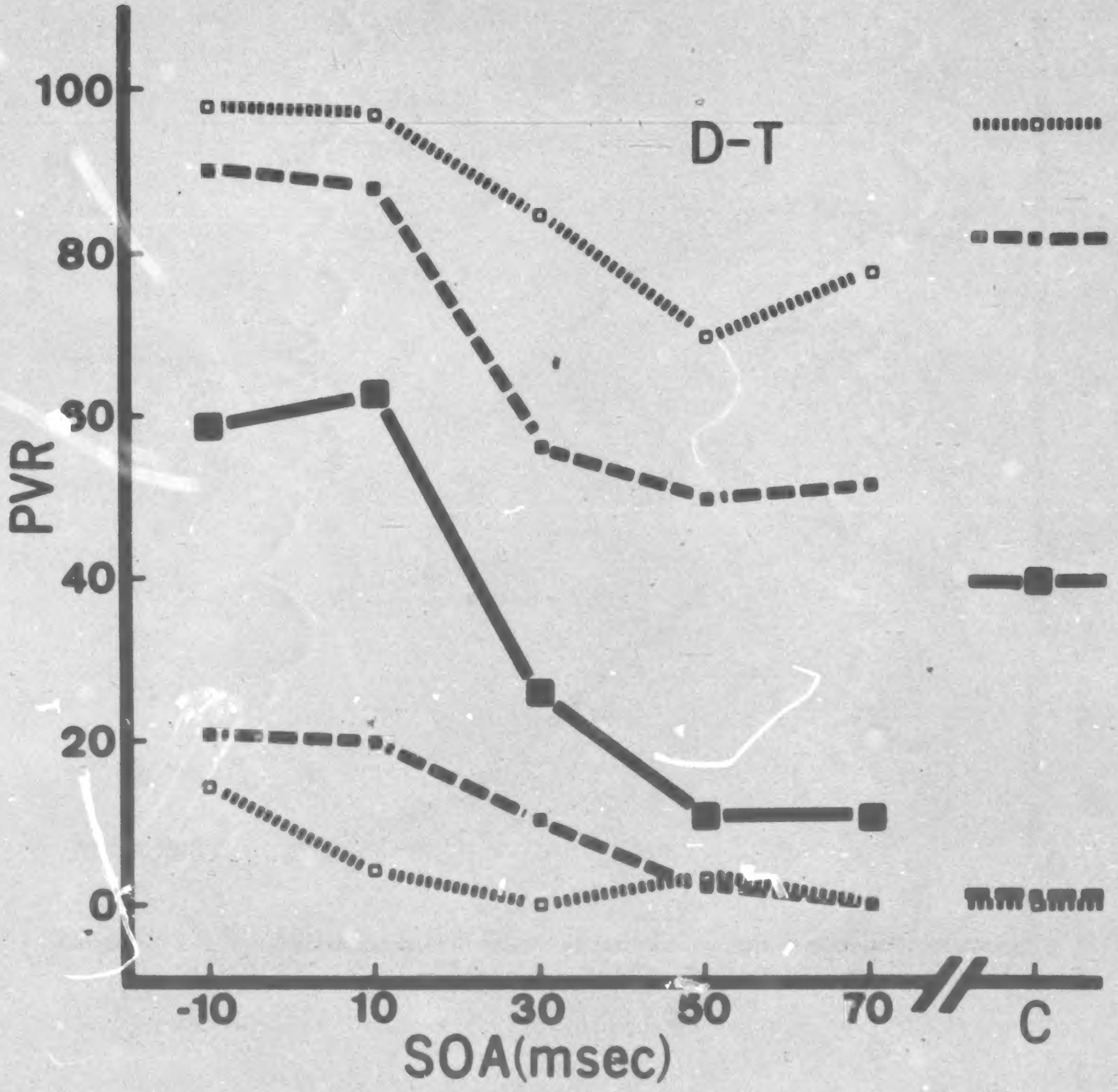

Figure 3: As Figure 2, for the D-T continuum.

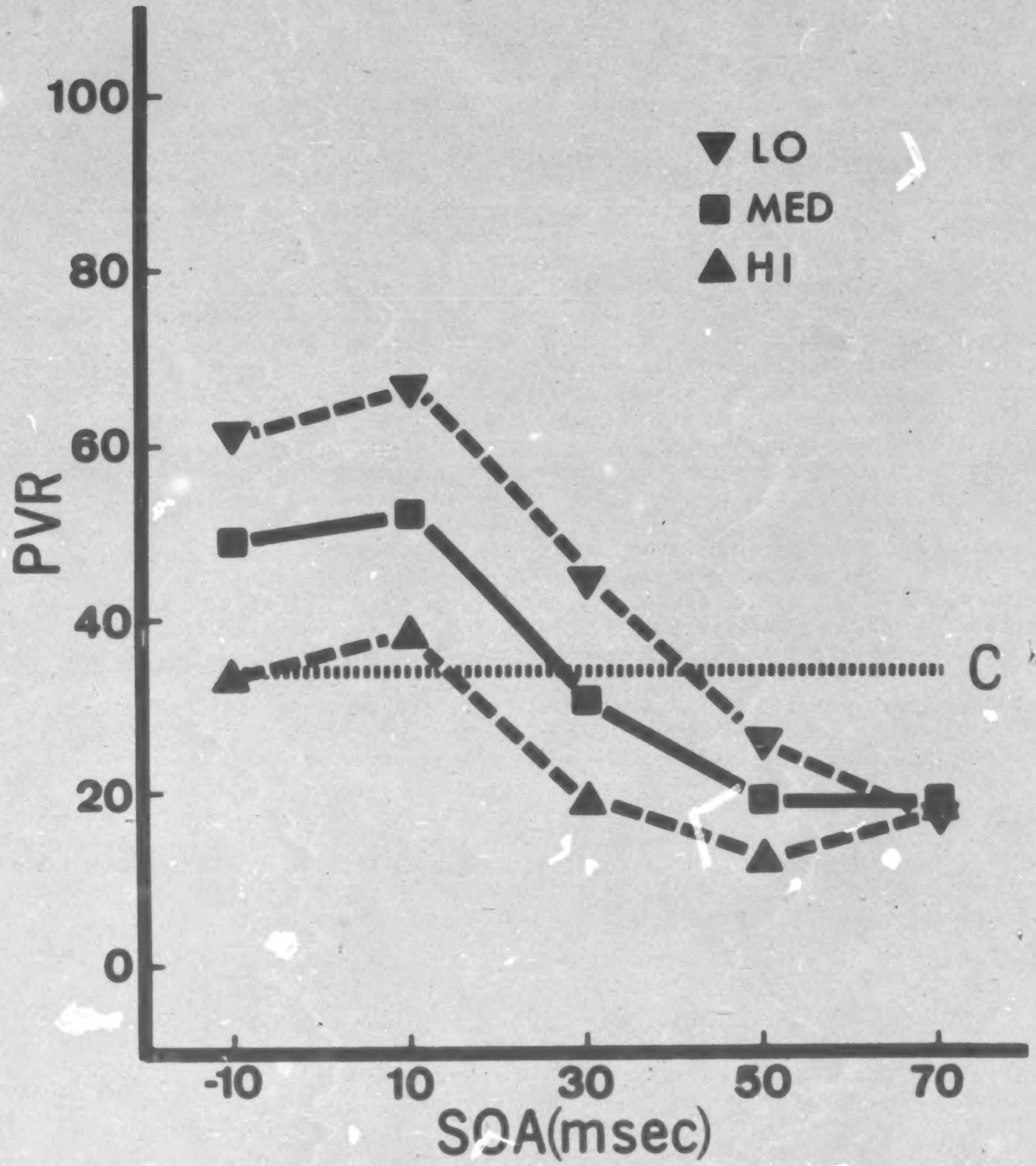

Figure 4: Average percentages of voiced responses (PVR) at five SOAs in the presence of dichotic vowel masks with three different fundamental frequencies at onset (LO, MED, HI). The average monaural PVR is shown here as a horizontal line (C). The data are weighted averages of all stimuli from both continua.

From Figure 4, it is also clear that the Pitch effect began to converge at SOA = 50 and suddenly disappeared completely at SOA = 70, although a negative bias was still present at this SOA. The Pitch × SOA interaction was highly significant ($p<.0008$) [for BHR, too ($p<.007$)]. The functions for BHR were remarkably similar to those in Figure 4, except that his difference between high and medium pitch disappeared by SOA = 30 and even tended to become inverted at the two longest SOAs.

Ear differences. Ear differences were expected to be manifested as an Ears × SOA interaction, such that the better ear showed both a smaller positive bias at short SOAs and a smaller negative bias at long SOAs. As shown in Figure 5, there was an indication of an average right-ear advantage (REA) but only at short SOAs, particularly SOA = -10. The interaction did not reach significance ($p>.24$). On the other hand, BHR showed a pronounced REA and a highly reliable interaction ($p<.0004$). For him, too, the REA was especially pronounced at SOA = -10.

The REA also tended to interact with Pitch. The ten subjects did not show any REA with high-pitch vowels, a tendency that approached significance ($p<.07$). The author showed a significant Ears × Pitch × SOA interaction ($p<.02$), due solely to the high-medium pitch contrast ($p<.005$): the REA was more pronounced at medium pitch at short SOAs, but at high pitch at long SOAs. The Place × Ears interaction also approached significance for BHR: his REA was more pronounced on the B-P continuum (at short SOAs only) than on the D-T continuum. Note that this last effect (more voiced responses to the left ear on the B-P continuum) is in agreement with a similar tendency observed in the monaural task.

Place errors. The ten subjects committed 7.8 percent place errors, which compares with 6.8 percent in monaural identification (or 3.8 percent, if only the second monaural series is considered). This is a relatively slight increase, and accuracy for place identification remained high. There were slightly more place errors in the right ear than in the left ear, but closer inspection showed an interaction with SOA: at the shortest SOA, there was a REA, while at the two longest SOAs, there was a left-ear advantage. This trend was nearly significant ($\chi^2_{(9)} = 15.3$, $.05<p<.10$). There were no other conspicuous patterns; note especially that place errors appeared to be equally frequent at all SOAs.

Summary of results. (1) When the target syllables are taken from a VOT continuum, isolated vowels in the other ear have a marked influence on the percentage of voiced responses. This percentage increases at SOAs shorter than about 30 msec and decreases at longer SOAs. The effect is most pronounced for target syllables close to the category boundary. Both the positive and the negative bias on voiced responses exceed the range of SOAs used here (-10 to 70 msec).

(2) The fundamental frequency at masking vowel onset systematically influences the percentage of voiced responses, higher frequencies leading to fewer voiced responses and lower frequencies leading to more voiced responses. This effect disappears quite abruptly at SOA = 70.

(3) While a group of naive subjects showed no significant REA, the results of BHR demonstrate that a REA may occur in this task. The REA tends to be most pronounced when the masking vowel leads in time.

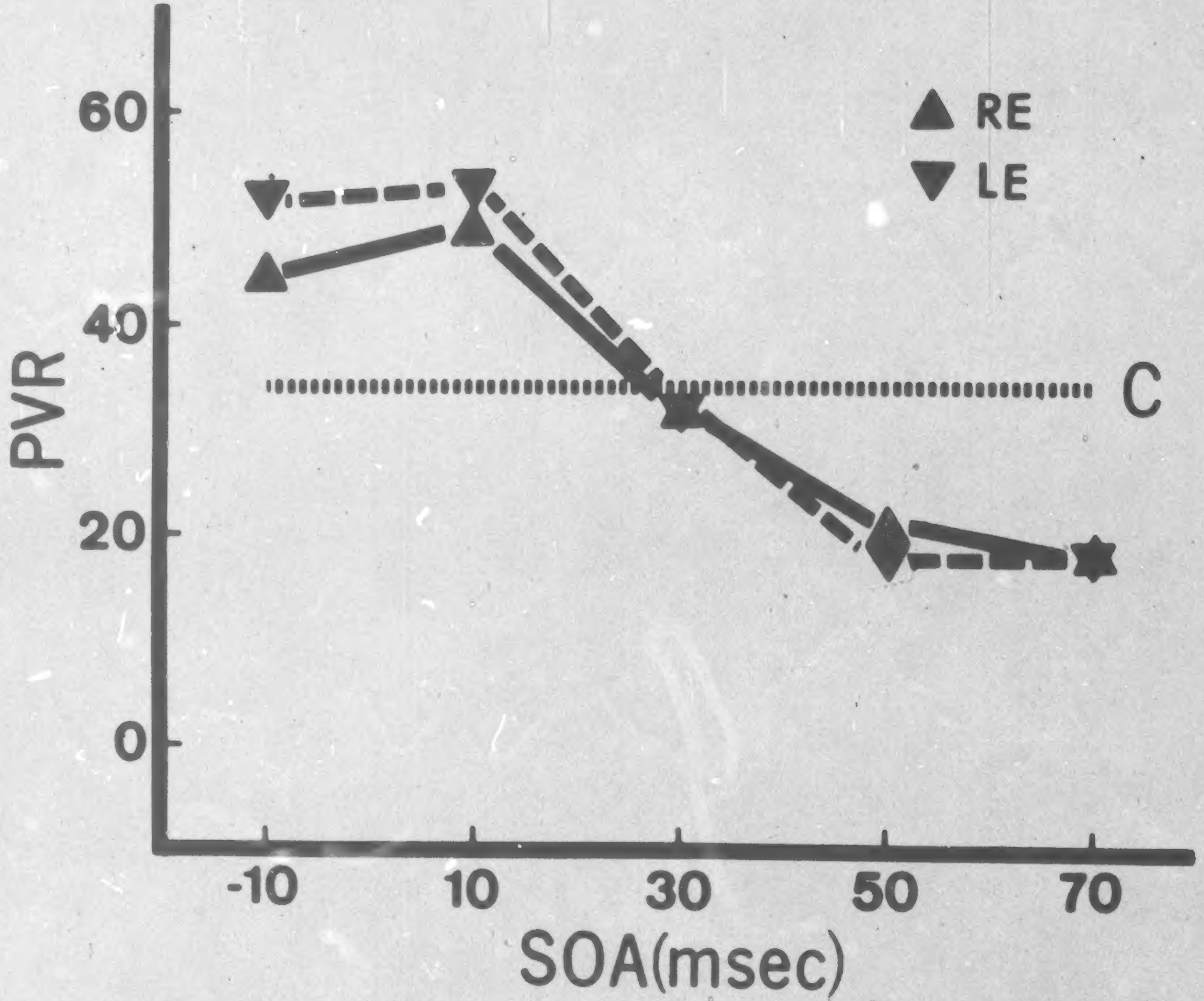

Figure 5: Average percentages of voiced responses (PVR) to the right ear (RE) and to the left ear (LE) at five SOAs in the presence of dichotic vowel masks. The average monaural PVR is shown as a horizontal line (C). The data are averaged over all other factors.

FIGURE 5

(4) Identification of the place feature is only little affected by the interfering vowels, and the place errors follow no clear pattern.

(5) All these effects apparently do not change with practice and (with the possible exception of the REA) are equally present in naive subjects and in an experienced listener.

EXPERIMENT II

Introduction

The second experiment was conducted for several reasons. First, there was the question whether the effects observed in Experiment I and in my previous study (Repp, 1975b) are specific to masking vowels that are phonemically identical with the vowel of the target syllables. To investigate this question, an identical vowel mask was compared with a phonemically very different vowel mask. Second, the range of SOAs was extended by 30 msec to both sides, in order to see whether the effects of the masking vowel would reach their asymptote within this range. Third, in an attempt to obtain more precise data in the region of the category boundary, the SOAs were spaced more narrowly in this critical region. The syllables were chosen from B-P and G-K continua, which are known to differ even more in their VOT category boundaries than labial and dental continua (Lisker and Abramson, 1967; Zlatin, 1974).

A fourth interesting aspect of the second experiment is its use of the vowel context /-i/. Summerfield and Haggard (1974) have demonstrated that VOT is of different salience as a voicing cue in /-a/ and /-i/ contexts. In /-a/ context, there is a substantial first-formant transition that, in conjunction with a delay in voicing onset, may act as a voicing cue ("transition detection" in the voiced part of the signal; cf. also Stevens and Klatt, 1974). To a lesser degree, this applies also to the /-ɛ/ context of Experiment I. In /-i/ context, on the other hand, the first-formant transition is minimal or absent, and VOT is likely to be the only cue to voicing (in the absence of a burst, as in the present syllables). On these grounds, the biasing effect of the "pseudo-VOT" (SOA) was expected to be even more pronounced in /-i/ context (targets of Experiment II) than in /-ɛ/ context (targets of Experiment I), since the transition of the first formant was more pronounced in the latter than in the former.

Other changes with respect to the first study included slightly higher intensities of both targets and masks, a short practice series at the beginning, and identical flat pitch contours for targets and masks (which permitted binaural fusion when the target and masking vowels were the same and overlapped). Pitch was not varied in this study.

Method

Subjects. Eight paid volunteer subjects participated. Some had participated in earlier experiments involving synthetic speech (but not in Experiment I) but all were more or less naive listeners. There were five females and three males. One male was left-handed. Again, BHR served as a comparison subject in six replications of the experiment.

Stimuli. This time, the target syllables were ten synthetic syllables from two VOT continua, /bi/-/pi/ and /gi/-/ki/. The VOTs of the five labials were

5, 15, 25, 35, and 45 msec, while those of the velars were 10 msec longer, viz. 15, 25, 35, 45, and 55 msec. The masking vowels were /i/ and /a/, each 250 msec long. Fundamental frequency was constant at 110 Hz for both targets and masks.

There were eight SOAs: -40, -10, 10, 20, 30, 40, 60, and 90 msec for the labial targets, and 10 msec longer for the velar targets, i.e., -30, 0, 20, 30, 40, 50, 70, and 100 msec. Target syllables from the middle of the continua were replicated three times and adjacent syllables twice. Hence, there were 2 (target continua) × 9 (5 VOTs—1 + 2 + 3 + 2 + 1 = 9 replications) × 2 (masking vowels) × 8 (SOAs) = 288 stimuli.

Procedure. Playback intensity was approximately 78 dB SPL for both targets and masks. Each subject first received a random practice series of 40 binaural syllables, which consisted of ten replications of each of the four "endpoint" stimuli of the two continua, that is, supposedly good instances of /bi/, /pi/, /gi/, and /ki/. The subject simply listened and compared the sounds with the correct responses on a list she had in hand. Subsequently, 144 monaural syllables were presented for identification, followed by the experimental series of 288 dichotic stimuli; and after a break the whole in reverse, with the target syllables in the opposite ear, just as in Experiment I. (All other details of method were the same as in Experiment I.)

Results

Monaural identification. Figure 6 shows the monaural voicing scores for the eight subjects. This time, the B-P continuum caused little trouble, but most subjects had considerable difficulties with the G-K continuum. One subject even gave inverted responses and tended to hear /g/ as /k/, and vice versa; her data are not included in Figure 1. Most subjects improved in the course of the experiment and did better on the second identification series (the functions in Figure 1 are based on both series), but the voicing identification function for G-K remained rather flat and did not reach asymptote at VOT = 55.

On the other hand, BHR gave equally consistent responses to both continua. In his data, there was a clear separation of category boundaries (which was indicated only at longer VOTs in the data of the eight subjects). This separation was about 15 msec, as compared to 6 msec in Experiment I, which is the expected outcome.

The eight subjects showed an average tendency to give more voiced responses to the left ear, but only on the G-K continuum, and due to only four listeners. However, BHR showed the same trend but much more consistently ($\chi^2(1) = 19.9$, $p < .0001$) on the G-K continuum, and to a lesser degree, also on the B-P continuum (as in Experiment I).

Place confusions were quite rare (2.2 percent) and occurred mostly on the B-P continuum. Thus, with the exception of one subject (the one with the inverted responses), the subjects' responses "stayed on the G-K continuum," despite the difficulties in voicing discrimination. BHR committed no place errors at all. No other patterns were observed in monaural place confusions, including absence of an ear difference.

The effect of SOA. The results are shown in Figures 7 and 8. In the statistical analysis, the eight SOAs were reduced to two levels by dividing the SOA

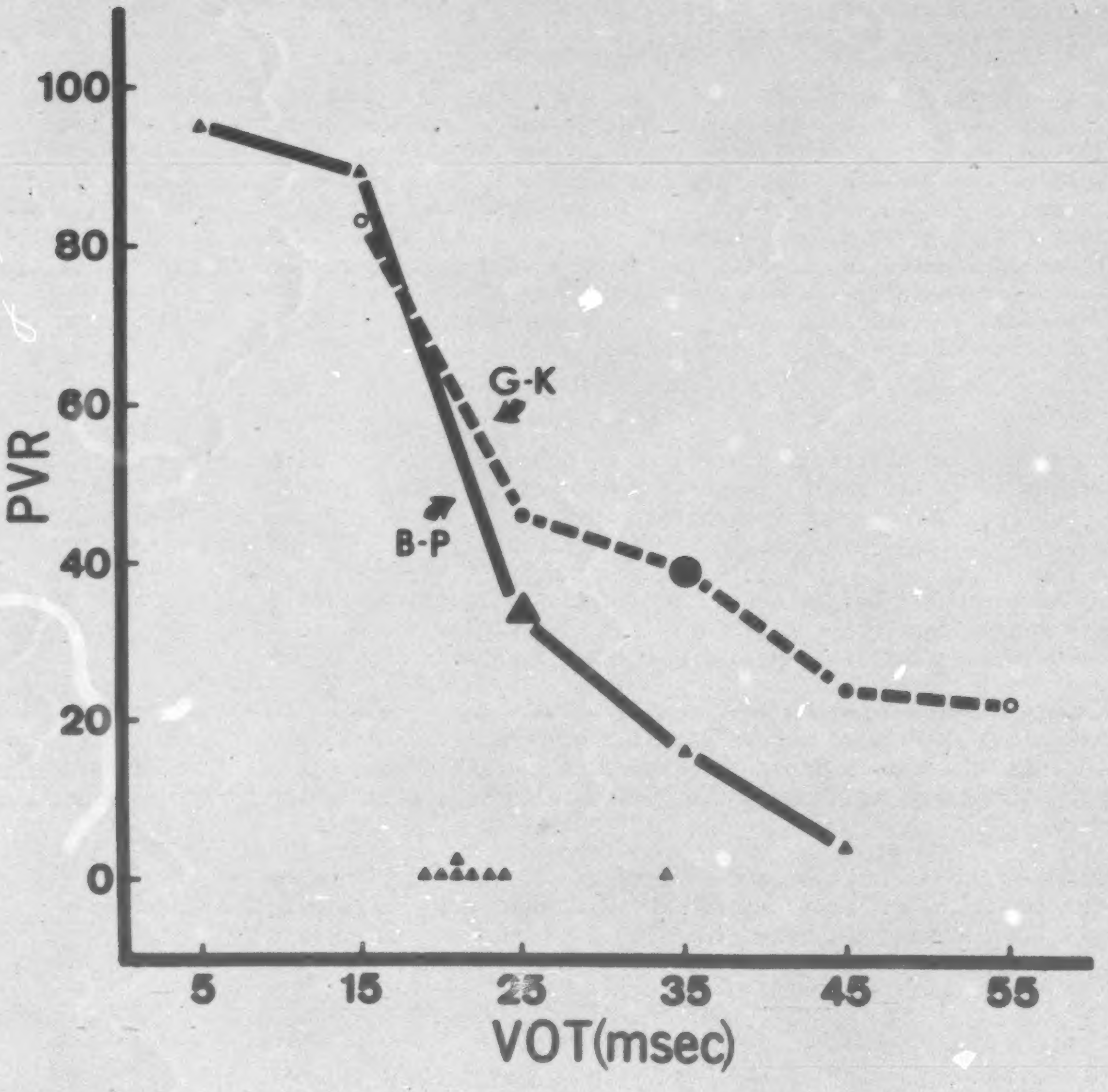

Figure 6: Average percentages of voiced responses (PVR) to monaural stimuli from labial and velar VOT continua in the context /-i/. The prominence of the individual data points (open - filled - large filled) is in proportion to the number of observations. The small triangles above the PVR = 0 level represent the distribution of the individual B-P category boundaries (PVR = 50). (The individual G-K boundaries are not shown because they were quite erratic.)

continuum into two halves and ignoring the 10-msec difference in SOAs for the two continua. (Strictly speaking, the *voicing onset asynchrony*, which was designed to be the same for the two continua, was substituted for SOA.)

As in Experiment I, SOA had the expected effect on the frequency of voiced responses. In fact, it was more dramatic than in Experiment I, as predicted, and was present even with the poor G-K continuum. The main effect of SOA (two levels) was highly significant ($p<.001$). Similarly for BHR ($p<.0001$). It interacted with the Place factor for the eight subjects only ($p<.02$), reflecting the relatively weaker effect of SOA on the G-K continuum, a direct consequence of the relatively poor voicing discrimination for these stimuli (Figure 8). Only BHR, on the other hand, showed a significant Place main effect ($p<.002$), due to a higher overall percentage of voiced responses on the G-K continuum. The fact that this expected main effect was not significant for the eight subjects is also in accord with the absence of a clear separation between the two identification functions (Figure 6), while BHR showed a much more pronounced separation.

Figures 7 and 8 show again that targets close to the category boundaries were affected most strongly, and that the relative extents of positive and negative biases depended on the baseline level. These effects were especially clear for BHR, while the results of the eight subjects showed some susceptibility to interference in all stimuli. The figures further demonstrate that the range of the SOA effect has not yet been bracketed by using SOAs between -40 and +100 msec: neither all positive nor all negative effects have reached their asymptotes at these SOAs. (See Figures 9 and 10 for better pictures of the average trends.)

The nature of the vowel mask. Figure 9 compares the average results for the two vowel masks. It is evident that /a/ did have a clear effect on voicing perception, but its effect was somewhat weaker than that of /i/. However, neither the Vowel main effect nor its interaction with SOA reached significance, although it is apparent that the difference between the two masks disappeared at the shortest and at the longest SOAs. (Both continua showed precisely the same pattern.) This interaction is not quite what would be expected if /a/ had had a truly weaker effect than /i/. On the other hand, BHR showed precisely the kind of interaction expected ($p<.002$), reflecting a genuinely attenuated but nevertheless present biasing effect of /a/.

Ear difference. Ear differences are shown in Figure 10. The pattern is strikingly similar to that in Experiment I (Figure 5). The eight subjects again showed a REA only in the region of positive voicing bias, and there was no clear interaction with SOA but instead a nonsignificant ($p<.10$) tendency to give more voiced responses to the left ear. Closer inspection of the data showed that the B-P continuum actually exhibited a weak interaction with SOA, while G-K showed only a strong main effect throughout (cf. the corresponding monaural tendency). The pattern in Figure 10 results from the averaging of these two effects. However, the triple interaction, Place × Ears × SOA, did not reach significance.

As in Experiment I, BHR showed the interaction expected from a true REA ($p<.008$). There was an interaction with Place ($p<.02$), reflecting a strong bias toward giving more voiced responses to the left ear on the G-K continuum (cf. the monaural data). The REA of BHR in the positive bias region was mainly due to G-K, while the REA in the negative bias region was due only to B-P. Given the higher frequency of voiced responses to the left ear in the monaural task, BHR's

FIGURE 7

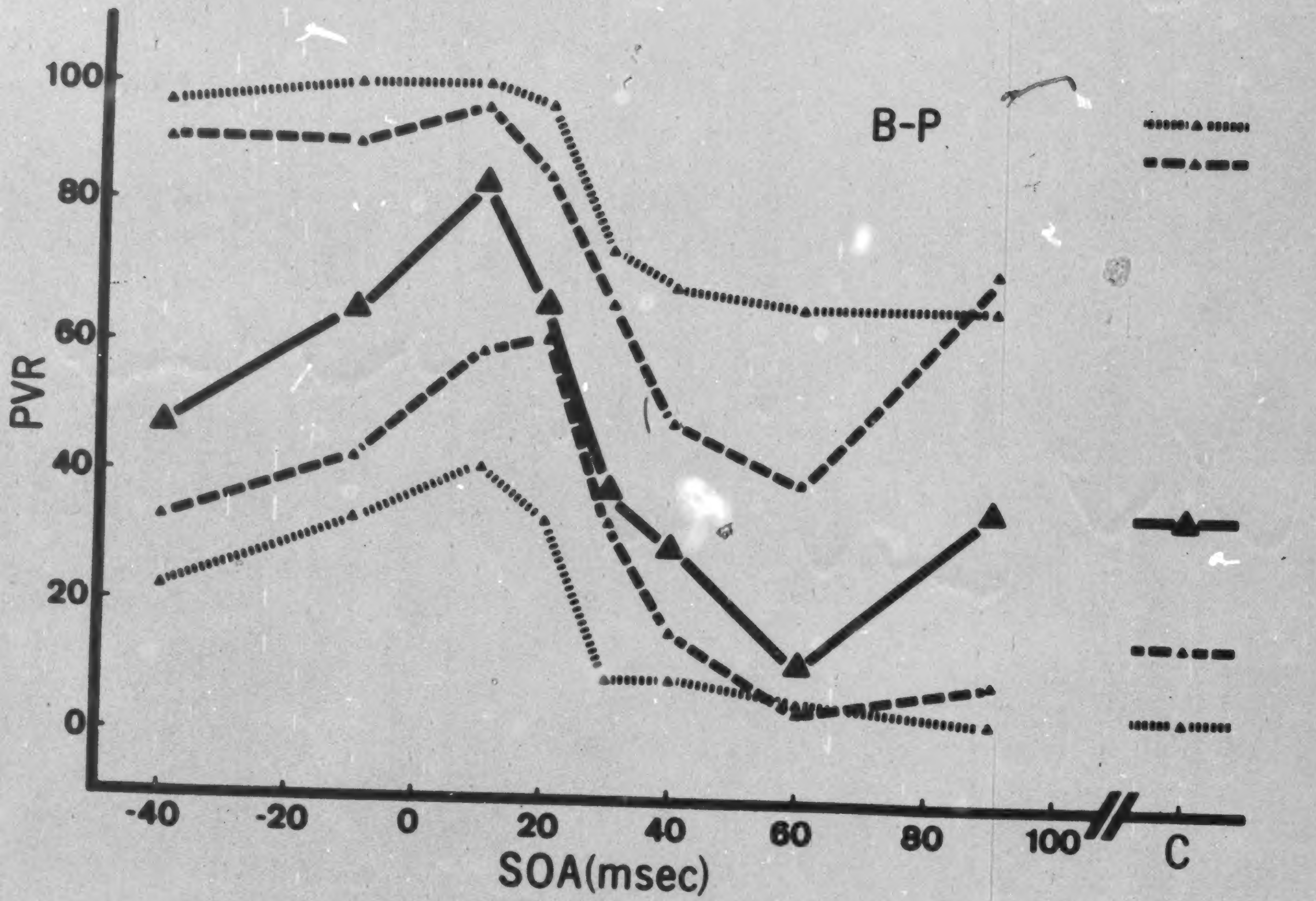

Figure 7: Average percentages of voiced responses (PVR) to the five stimuli from the B-P continuum, in the presence of vowel masks in the other ear at eight different SOAs. The prominence of the individual data points and of their connecting lines is in proportion to the number of observations. The monaural control PVRs (from Figure 6) are shown at the right (C). Large symbols represent VOT = 25; small filled symbols, VOT = 15 (top) and 35 (bottom); and small open symbols, VOT = 5 (top) and 45 (bottom). The data are averaged over the Vowel and Ears factors.

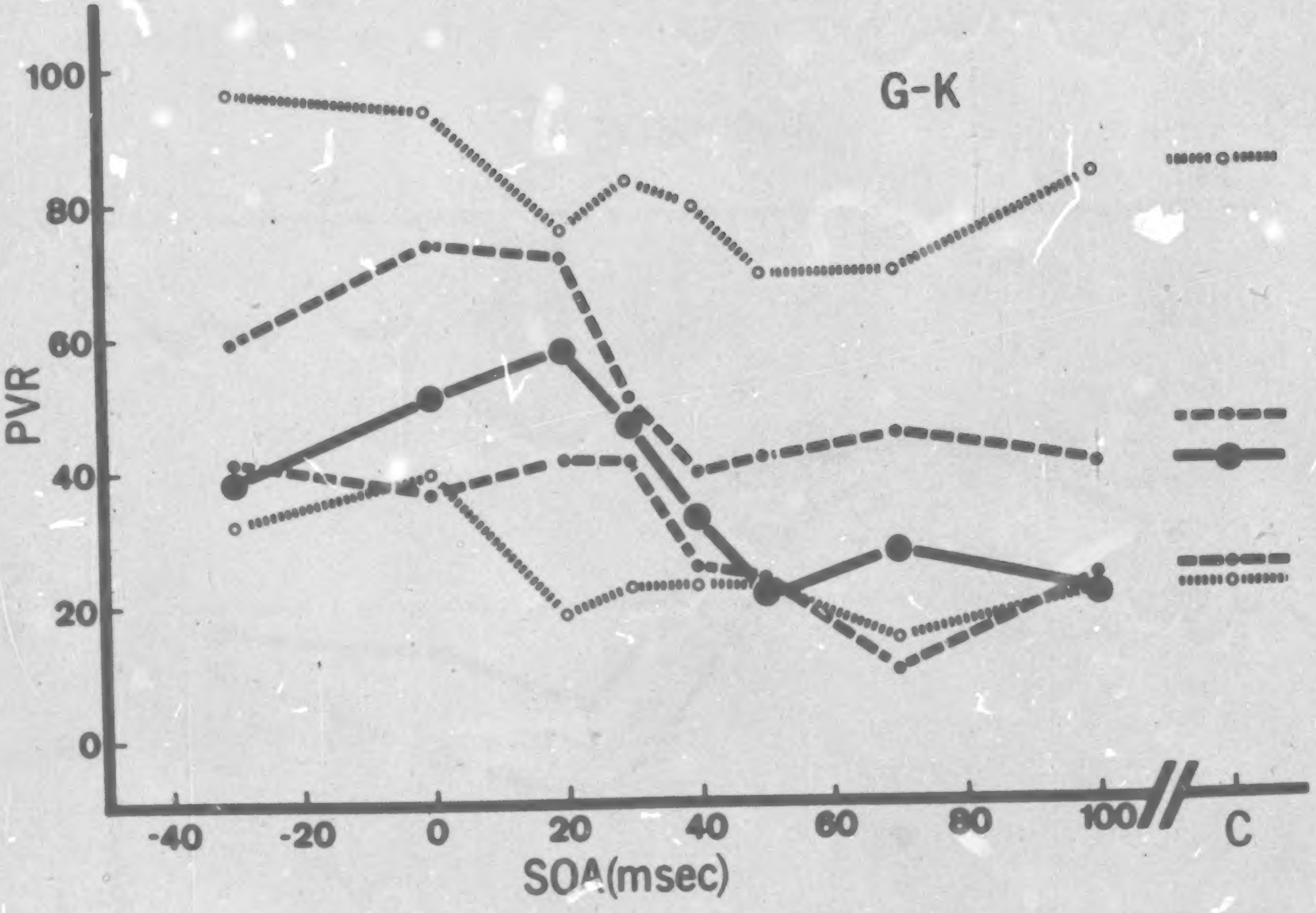

Figure 8: As Figure 7, for the G-K continuum. The VOTs are 10 msec longer here.

FIGURE 8

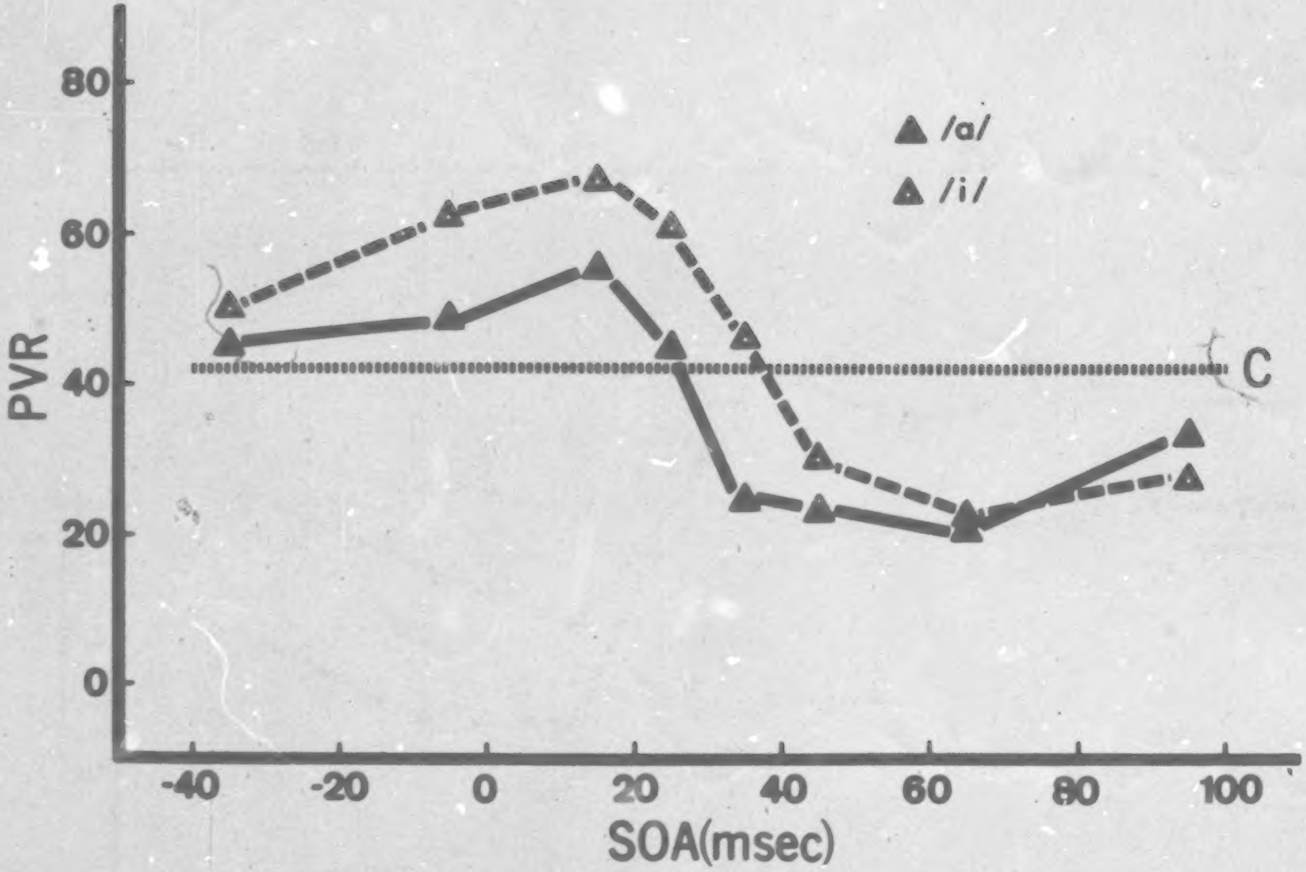

Figure 9: Average percentages of voiced responses (PVR) at eight SOAs in the presence of two different vowel masks (/a/ and /i/) in the other ear. The average monaural PVR is shown as a horizontal line (C). The data are weighted averages of all stimuli, and averaged results from the two continua are plotted at the average of corresponding SOAs (e.g., the points at SOA = -35 are the averages of the B-P data at SOA = -40 and the G-K data at SOA = -30, etc.).

FIGURE 9

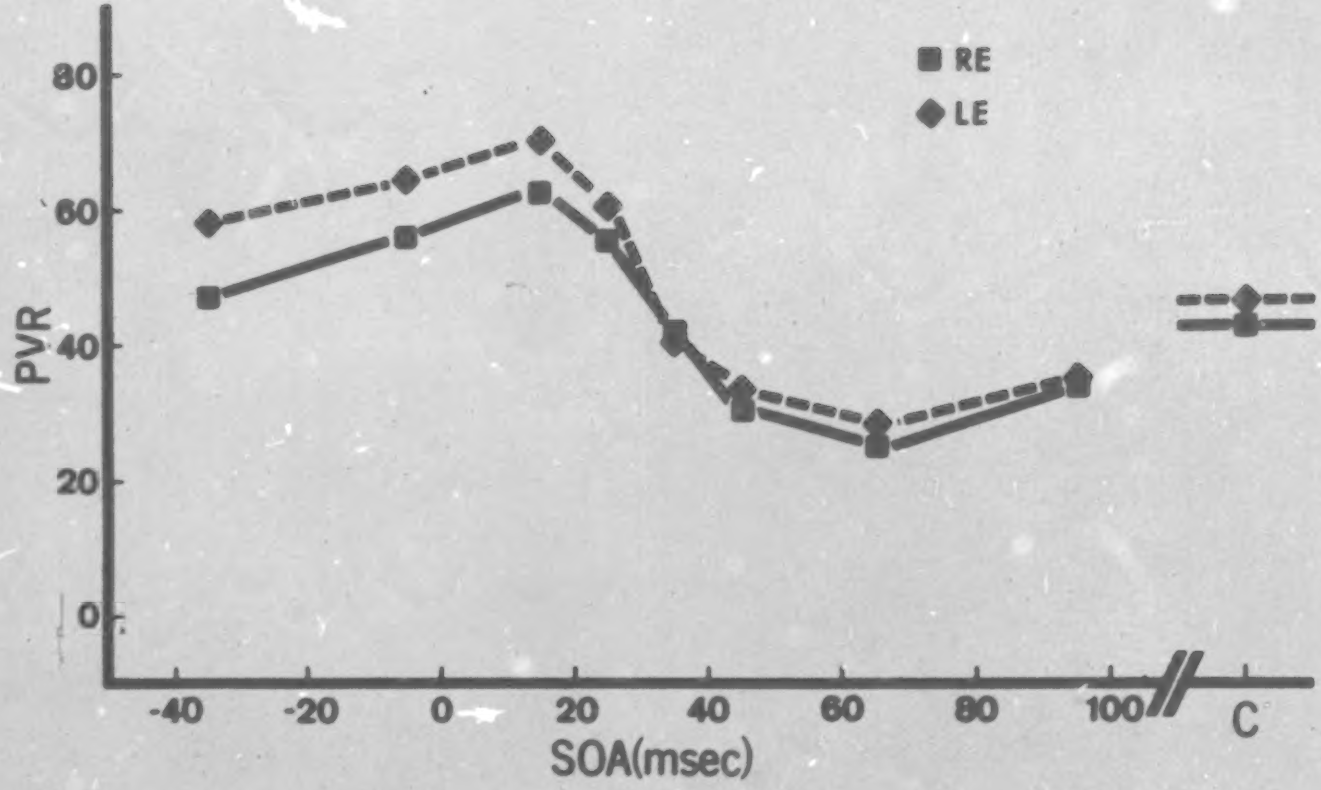

Figure 10: Average percentages of voiced responses (PVR) to the right ear (RE) and to the left ear (LE) at eight SOAs (averaged between continua) in the presence of dichotic vowel masks. The average monaural PVRs are indicated separately for the two ears at the right (C). The data are averaged over all other factors.

FIGURE 10

results may also be described as showing approximately equal positive voicing bias in both ears, but a negative bias only in the left ear and not at all in the right ear. His average REA was most pronounced at SOA = -40.

Place errors. The place confusions committed by the eight naive subjects increased from 2.2 percent (monaural) to 3.3 percent (dichotic), again only a small increase. However, the errors showed a definite pattern: they were much more frequent with /i/ as mask than with /a/ (5.0 percent vs. 1.6 percent; $\chi^2(1) = 45.2$, $p<.0001$). In other words, only the /i/ mask led to an increase in place errors at all. Place confusions were also more frequent in the left ear than in the right ear (4.0 percent vs. 2.6 percent; $\chi^2(1) = 5.0$, $p<.03$). No clear pattern with respect to the target stimuli or SOA was present, except that errors tended to be more frequent in the G and P regions. (A similar trend existed monaurally.) BHR committed only five place errors (0.1 percent), which--perhaps not by coincidence--all occurred in the left ear and with /i/ as mask.

Summary of results. (1) The basic effect of vowel masks on the perception of voicing was replicated with a different set of stimuli in the context /-i/. The effect was even more striking than in Experiment I and still exceeded the boundaries of the SOA range (-40 to 100 msec).

(2) The effects of a vowel mask identical with the vowel following the target consonant (/i/) and of a very different vowel mask (/a/) were qualitatively similar, but the effects of /a/ tended to be weaker.

(3) Again, BHR showed a clear REA, while the group of eight subjects showed only a tendency in that direction. Also, the REA was again most pronounced at negative SOAs.

(4) Place confusions were increased only with the /i/ mask and tended to exhibit a REA.

DISCUSSION

The Nature of the "VOT Masking Effect"

The present experiments confirm and extend the observations of Repp (1975b) by demonstrating a clear biasing effect of isolated vowel masks on the perception of the voicing feature of a stop-consonant target in the other ear. This effect is very different from the kind of interference observed in typical dichotic listening tasks, where two equivalent stimuli (e.g., CV syllables) compete for a single processor or response mechanism (Pisoni and McNabb, 1974; Repp, 1974a). Given consonants (in CV syllables) as targets, isolated vowel masks contain no competing auditory or phonetic information that could be mistakenly accepted as a consonantal feature. Rather, it is the temporal relationship between target and mask that itself provides competing information. Thus, in contrast to a standard masking paradigm, the competing information changes with SOA. Of course, the relevant aspect of the target, the voicing feature, is also defined primarily as a temporal relationship, VOT, which was the only voicing cue in the synthetic syllables used here. In a way, then, a temporal relationship is being "masked" by another temporal relationship in the present paradigm.

This explanation in terms of derived auditory or phonetic parameters implies that the interaction between the competing information takes place after temporal relationships have been extracted from (and from between) the auditory signals, that is, at a higher auditory or phonetic level. This is in agreement with a general explanation of dichotic interference in terms of interactions of derived stimulus features (e.g., Blumstein, 1974). It is conceivable, however, that similar effects arise from the combination of more elementary auditory information from the two ears. The two stimuli could fuse with each other and be acoustically integrated before temporal parameters are extracted from this combined signal, which would represent a complex case of simultaneous "central" auditory masking (Repp, 1975b). There are indications, however, that a higher-level interaction is involved. One hint is that the perception of the place feature is so little affected. If the masking vowel were truly superimposed on the target stimulus, it should have a stronger simultaneous masking effect on the transition portion of the target syllable whenever the two overlap in time. However, place errors showed no clear trends with respect to SOA. A second argument is the relatively wide temporal range of the voicing bias, which is estimated to reach from about -70 to 120 msec SOA. This range by itself points toward a higher level of interaction. Perhaps the strongest argument concerns the reduced effects of vowel masks on target stimuli that lie farther away from the VOT category boundaries. This is best illustrated by considering two simple models of the "VOT masking effect."

Assume that, in each target-mask combination, there are two competing voicing cues: the actual VOT of the target syllable, and the "pseudo-VOT" due to the SOA between target and mask onsets. If these two cues were simply weighted and combined (or, in terms of a discrete model, if one were substituted for the other in a certain percentage of cases), the resulting average percept should lie somewhere between the two extremes corresponding to the two cues. This can be expressed as

$$p(V+ \mid VOT, SOA) = (1 - c_{SOA})p(V+ \mid VOT) + c_{SOA}p(V+ \mid SOA) \qquad (1)$$

which is to be read as follows: the probability of a voiced response to a target with a given VOT at a given target-mask SOA, $p(V+ \mid VOT, SOA)$, is a weighted combination of the probability of a voiced response for this target in isolation, $p(V+ \mid VOT)$, and the probability of a voiced response given only the SOA ("pseudo-VOT") cue, $p(V+ \mid SOA)$. The c_{SOA} factor is a measure of the relative influence of the SOA cue; it is 0 when there is no influence and 1 when $p(V+ \mid VOT, SOA)$ depends on SOA only. The reason c_{SOA} depends on SOA is that the influence of the vowel masks clearly diminishes as SOA increases in either direction.

One can easily estimate $p(V+ \mid VOT)$ by the identification function for isolated syllables. On the other hand, $p(V+ \mid SOA)$ is unknown and probably could not be measured by itself at all. However, if SOA acted indeed like a "pseudo-VOT," the $p(V+ \mid SOA)$ should follow a similar function as $p(V+ \mid VOT)$, with the "category boundary" lying around the SOA that equals the VOT at the category boundary. However, while the VOT category boundary depends on place of articulation, the SOA category boundary may well be independent of this within-syllable factor.

The model in Equation (1) can easily be tested. It states that, at any given SOA, $p(V+ \mid VOT, SOA)$ should be a linear function of $p(V+ \mid VOT)$, with the slope $(1 - c_{SOA})$ and the intercept $c_{SOA}p(V+ \mid SOA)$. Plotting $p(V+ \mid VOT, SOA)$ as a function of $p(V+ \mid VOT)$ at different SOAs is an instructive alternative way of summarizing the data. These functions are shown in Figures 11 and 12 for the two

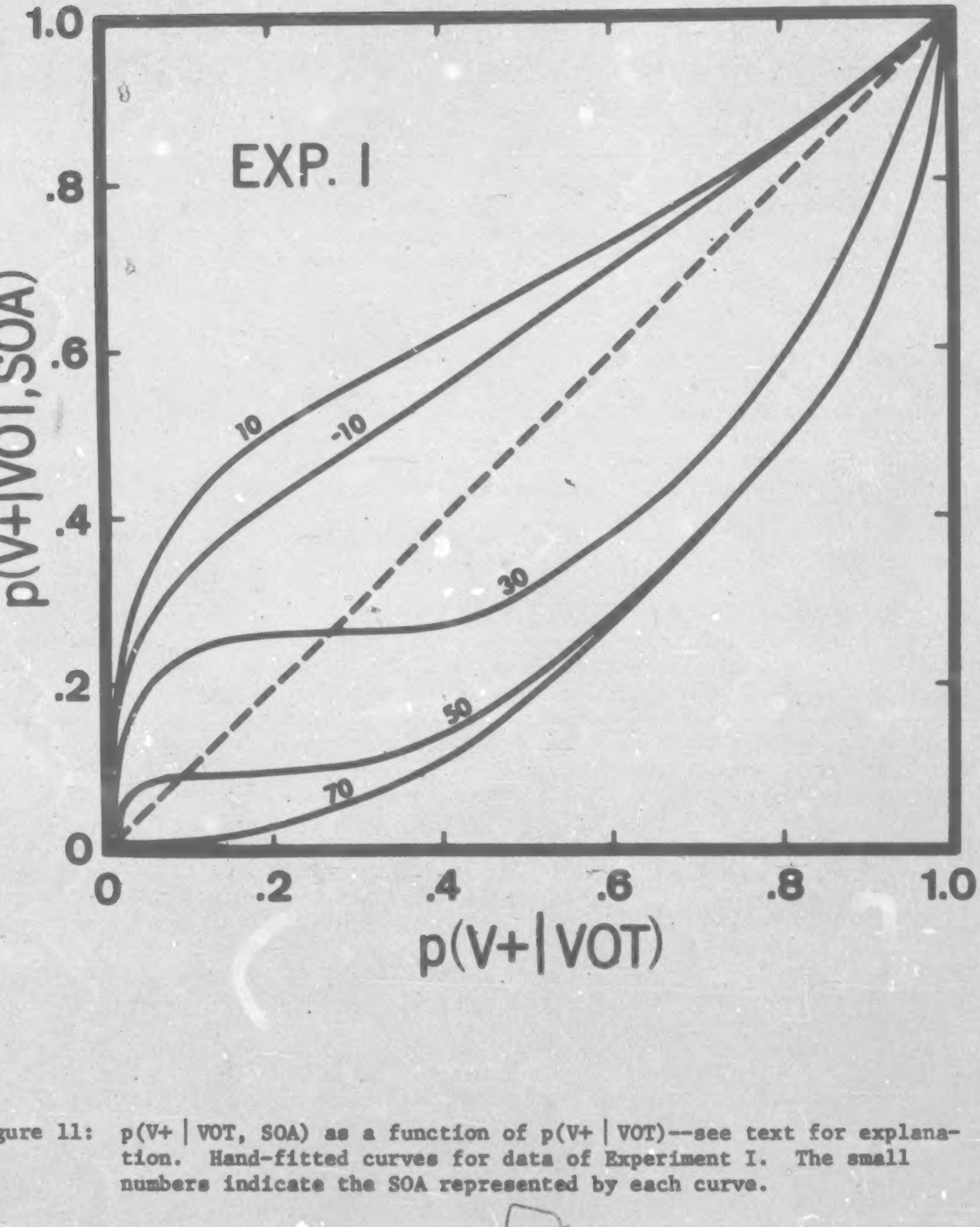

Figure 11: p(V+ | VOT, SOA) as a function of p(V+ | VOT)--see text for explanation. Hand-fitted curves for data of Experiment I. The small numbers indicate the SOA represented by each curve.

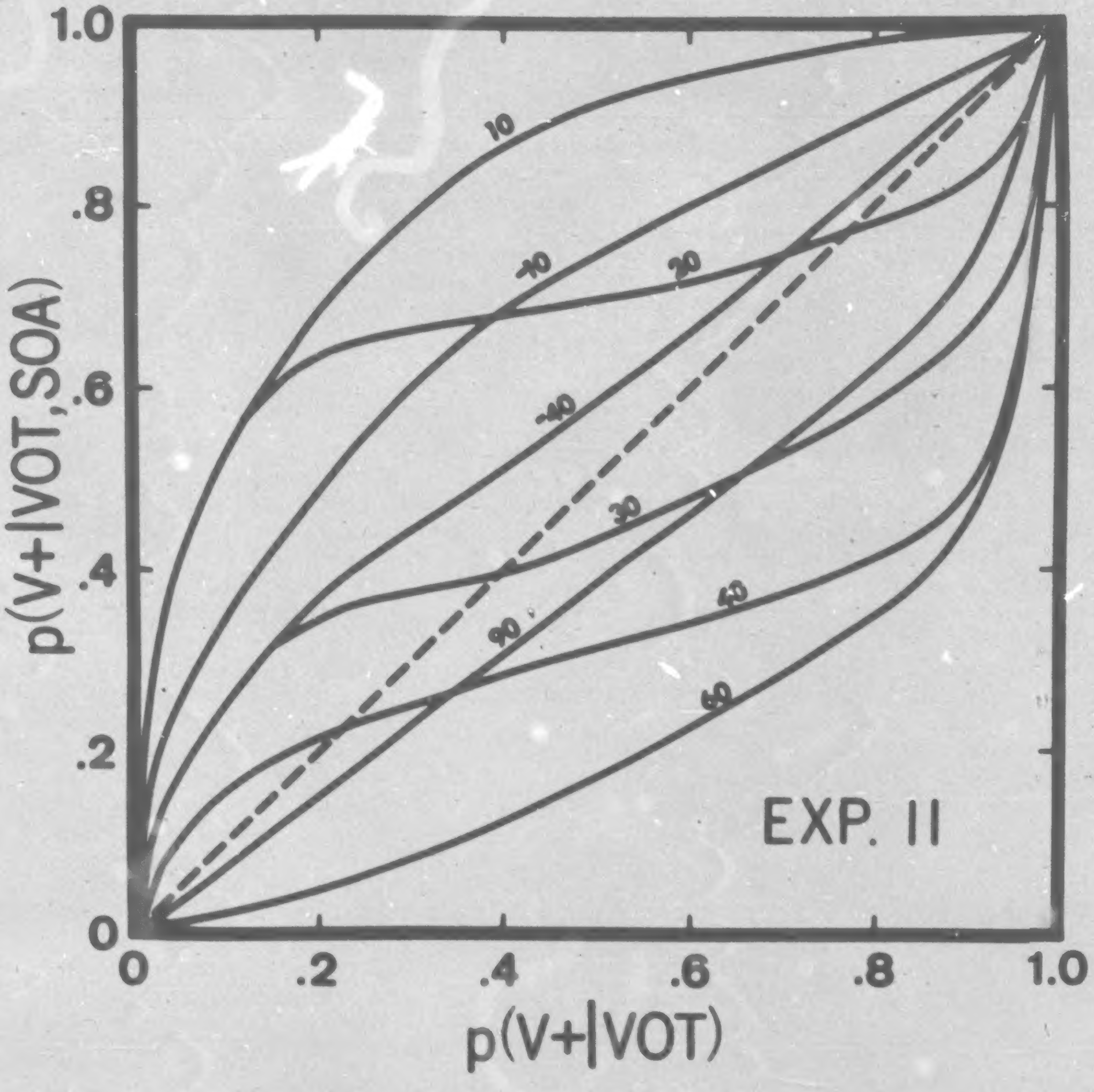

Figure 12: As Figure 11; data from Experiment II.

experiments, respectively. Figure 11 is derived from Figures 2 and 3. The curves were fitted by hand to the 10 points corresponding to the 10 target stimuli. Apparently, both B-P and D-T stimuli were fitted well by the same function, as predicted. In other words, the effect of SOA seems to depend only on $p(V+ \mid VOT)$ but not on place of articulation (as far as this rough analysis goes). Figure 12 represents the B-P stimuli from Figure 7. The G-K stimuli were not fitted well by these functions and were omitted; their erratic pattern was probably due to the perceptual difficulties the subjects had with them.

The functions for the two experiments are essentially similar, and all are grossly nonlinear. (BHR showed very similar patterns.) The simple model of Equation (1) must therefore be rejected. The relationship between $p(V+ \mid VOT, SOA)$ and $p(V+ \mid VOT)$ can be described in terms of a family of curvilinear functions that is characterized by two parameters: (1) the amount of (maximal) deviation from the $y = x$ (no effect) diagonal, which represents the degree of influence of SOA; and (2) the point at which the function crosses the diagonal, that is, the $p(V+ \mid VOT)$ at which a particular SOA has no effect. (This point is 0 if every SOA exerts a negative bias, and 1 if every SOA exerts a positive bias.) The precise mathematical representation of these functions still needs to be derived.

Note that it is not true that SOA has no effect when $SOA = VOT$, as one might expect intuitively (cf. Figures 2, 3, 7, and 8). Rather, the "neutral" point depends on SOA and on $p(V+ \mid VOT)$ but not directly on VOT. It is then straightforward to make the assumption that $p(V+ \mid VOT, SOA) = p(V+ \mid VOT)$ precisely when $p(V+ \mid VOT) = p(V+ \mid SOA)$; that is, that SOA will have no biasing effect when the probabilities of a voiced response on the basis of each cue are equal [cf. Equation (1)]. Given this assumption, the hypothetical $p(V+ \mid SOA)$ functions can be derived from Figures 11 and 12, and they are plotted in Figure 13. Despite this very rough estimation procedure, the functions derived from the two experiments are remarkably similar, as they should be, although the average SOA effects were more pronounced in Experiment II. Both functions resemble VOT identification functions (cf. Figures 1 and 6, the better stimuli) and have category boundaries [where $p(V+ \mid SOA) = .5$] at 24- and 28-msec SOA, respectively. These values are in the same range as the VOT boundaries for the stimuli used here. The separation of the two functions in Figure 13 may be due to various factors (vowel contexts, intensities, pitch contours, etc., or merely imprecision) and deserves no further comment at this stage. (For the sake of fairness, it should be mentioned that the corresponding functions for BHR were not nearly as neat and had atypical category boundaries. Therefore, some scepticism about this analysis remains.)

Clearly, SOA has less of an effect on stimuli remote from the VOT category boundary [whose $p(V+ \mid VOT)$ is close to 0 or 1] than on stimuli close to the boundary. One reason may be that the effect comes about through "direct" masking of the voicing onset in the target by the (voicing) onset of the mask, which might be a relatively peripheral process, as suggested by Repp (1975b). In this case, the voicing onset asynchrony (VOA) would determine the extent of the bias. The model of Equation (1) may then be reformulated:

$$p(V+ \mid VOT, SOA) = (1 - c_{VOA})p(V+ \mid VOT) + c_{VOA}p(V+ \mid SOA) \qquad (2)$$

The weighting factor now depends on $VOA = SOA - VOT$. This model can be tested if $p(V+ \mid SOA)$ is estimated by the functions in Figure 13. However, an attempt to

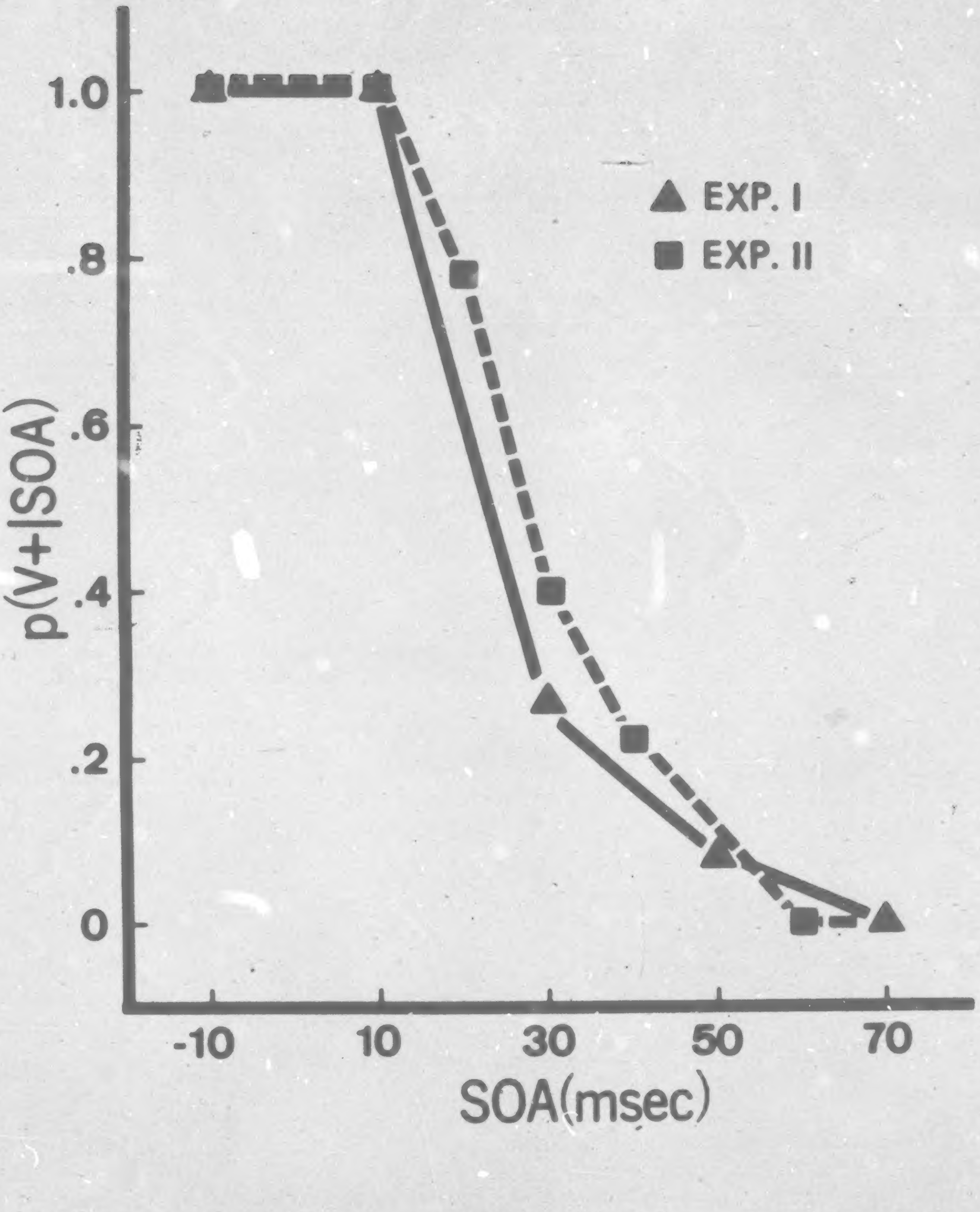

Figure 13: p(V+ | SOA) as estimated from Figures 11 and 12--see text for explanation.

fit this model failed: the fit was excellent in the middle range of VOT, but the effect of SOA on stimuli with extreme p(V+ | VOT) was grossly overestimated. Hence, Equation (2) must be rejected, too.

These results confirm in a somewhat more precise fashion what Figures 2, 3, and 7 seemed to show: the effect of SOA depends on p(V+ | VOT). Stimuli that have a low a priori "voicing uncertainty" are less affected than stimuli with high voicing uncertainty, even taking into account factors such as VOA. This seems to provide a fairly strong argument that the effect takes place at a relatively late stage in processing, vis. at the linguistic level where voicing decisions are made. It seems that an attempt to categorize (the voicing dimension of) the target stimulus precedes the influence of SOA, and that the degree of this influence then depends on whether the categorization has had an ambiguous outcome or not. The complexities of deriving a formal model for this case will be avoided here, especially since inspection of Figures 11 and 12 indicates that maximal SOA effects do not always occur with stimuli of maximal voicing uncertainty [p(V+ | VOT) = .5] but depend in a more complex fashion on both uncertainty and the "room for variation" in a certain direction.

It is interesting to note that the negative bias tended to be stronger and had a wider temporal range than the positive bias. This difference is analogous to the "lag effect" observed in traditional dichotic paradigms (Studdert-Kennedy, Shankweiler, and Schulman, 1970; Repp, 1975a). Apparently, at the level of competing temporal relationships, too, the lagging information tends to displace the leading information more often than vice versa. This is another argument supporting the conclusion that the vowel "masking" effect arises at a relatively late stage in processing.

Secondary Effects

The effects of SOA were more pronounced in Experiment II than in Experiment I, especially on the B-P continuum. Although there were several other differences between the two experiments (for example, intensity levels, pitch contours, subjects), the most appealing explanation is with regard to the vowel context, as outlined earlier. In the /-ε/ context, the subjects presumably could rely on an additional cue to voicing (the first-formant transition), besides the temporal aspect of VOT, and therefore were somewhat less susceptible to the effects of SOA, especially since the masking vowels hardly seemed to affect transitional information (as shown by the small increase in place errors).

Experiment II further demonstrated that the biasing effect is obtained with a masking vowel (/a/) that is very different from the vocalic context of the target (/-i/). Since the effect of /a/ was reduced relative to that of /i/, it seems that the "perceptual separation" of the dichotic stimuli influences the extent of "masking" to some degree. Especially with the flat pitch contours used in Experiment II, identical vowels in the two ears tended to fuse when they overlapped, resulting in a partially fused percept. This did not occur with a very different masking vowel, so that it could be more easily ignored. However, since a biasing effect was obtained even with /a/ masks, the most important factor seems to be the onset of energy in the region of the fundamental frequency. It is interesting to note, in addition, that the /i/ mask increased place errors while the /a/ mask did not. For the perception of transitions in the target, the formant structure of the mask may well be more relevant than for the perception of VOT.

The effects of fundamental frequency variation were investigated in Experiment I. Three vowel masks that all terminated at the same pitch had strikingly different effects, depending on their starting frequency. The lower the fundamental, the stronger the tendency to give voiced responses. Clearly, this effect is related to "pitch as a voicing cue," as described by Haggard, Ambler, and Callow (1969). However, the stimuli employed by Haggard and his colleagues used much more extreme "pitch skips" within synthetic syllables, which occurred during the transitions, while the present differences were smaller and, of course, occurred in the opposite ear. Perhaps more comparable is the finding (House and Fairbanks, 1953) that, in natural speech, fundamental frequency tends to be higher following voiceless stops than following voiced stops, and that this difference extends into the following vowel. The rapid convergence of the pitch effect at SOA = 70 in Experiment I is interesting; it would be challenging to search for evidence that pitch differences tend to extend for a comparable time into the steady-state vowel in natural speech. Haggard's estimate[1] of about three pitch periods (about 30 msec) is precisely of the right magnitude. In any case, the pitch effect may be added to the growing list of examples that speech perception takes into account the constraints and dynamics of speech production.

The final issue to be discussed is the right-ear advantage. For the naive subjects, it did not reach significance in either experiment, but there was a tendency that was strikingly similar in both studies in that it was strongest at negative SOAs. The author, who is known to exhibit a pronounced REA in a standard dichotic listening test, showed highly significant REAs in both experiments. He also showed the strongest REA at negative SOAs. His results are accepted as sufficient evidence that individual ear advantages may manifest themselves in this task, although the data of the naive subjects suggest that the REA is less pronounced than in dichotic tests employing competing CV syllables. Following current theories of lateralization, the REA may be assumed to be due to transcallosal degradation of left-ear stimuli in the presence of competing right-ear input (cf. Studdert-Kennedy, 1975). Apparently, an isolated vowel mask is sufficient to provide the competition necessary to inhibit the ipsilateral auditory pathways, although it does not compete directly at the phonetic level. This is reasonable, since this kind of inhibition is certainly governed by gross auditory characteristics. Transcallosal degradation may lead to less accurate identification of the voicing cues in the target. In addition, the left hemisphere may be more precise in assessing brief time intervals such as VOT, but because of their linguistic nature, left-hemisphere processing of the target syllables is highly likely. "Pseudo-VOT," on the other hand, is a relation between the two ears, and its perception should not be subject to laterality differences. The pitch at masking vowel onset is a property that need not necessarily lead to a REA; in fact, prosodic variables have often been associated with a left-ear superiority. It is interesting to note that neither BHR nor the subjects showed any tendency toward an ear difference for the Pitch effect in Experiment I. Only the effect of SOA was associated with a REA. Why this REA was most pronounced at negative SOAs is not quite clear, but the phenomenon cannot be ignored; it was too consistent. Perhaps, ipsilateral inhibition is maximal when there is complete overlap of the dichotic stimuli; or inhibition takes time to develop and therefore is maximal when mask onset precedes target onset. However, there is no corresponding phenomenon in conventional dichotic listening experiments.

---

[1] Mark Haggard, 1975: personal communication.

While the REA in voicing perception did not reach significance for naive subjects, they showed a significant REA in place perception in Experiment II and at the shortest SOA in Experiment I. (BHR made too few errors to show any significant ear differences.) This is in line with the classical REA in dichotic listening. It is interesting that the weak interference provided by the vowel was sufficient to reveal a REA. This REA may be a lower-level component of the REA obtained when the mask is a CV syllable, since, in the present case, the competing phonetic information on the place dimension is eliminated (and with it a corresponding part of auditory competition, otherwise provided by the transitions of the mask). This lower-level component may represent transcallosal degradation and/or auditory interference only, while competing CVs may add an additional REA owing to phonetic competition and transition-specific auditory competition. Such an interpretation fits well the increasing awareness that ear advantages may be composed of processing superiorities at several levels (Porter and Berlin, 1975).

The ear differences exhibited in the monaural VOT boundaries by BHR remain a curious and puzzling effect. It deserves some further investigation, since it suggests the possibility that monaural ear differences may exist in the perception of temporal stimulus properties.

REFERENCES

Blumstein, S. E. (1974) The use and theoretical implications of the dichotic technique for investigating distinctive features. Brain Lang. 1, 337-350.

Bock, R. D. (1975) Multivariate Statistical Methods in Behavioral Research. (New York: McGraw-Hill).

Darwin, C. J. (1971) Ear differences in the recall of fricatives and vowels. Quart. J. Exp. Psychol. 23, 46-62.

Efron, R. (1963) The effect of handedness on the perception of simultaneity and temporal order. Brain 86, 261-284.

Haggard, M., S. Ambler, and M. Callow. (1969) Pitch as a voicing cue. J. Acoust. Soc. Amer. 47, 613-617.

Halperin, Y., I. Nachshon, and A. Carmon. (1973) Shift of ear superiority in dichotic listening to temporally patterned nonverbal stimuli. J. Acoust. Soc. Amer. 53, 46-50.

House, A. S. and G. Fairbanks. (1953) The influence of consonant environment upon the secondary acoustical characteristics of vowels. J. Acoust. Soc. Amer. 25, 105-113.

Lisker, L. and A. S. Abramson. (1967) The voicing dimension: Some experiments in comparative phonetics. In Proceedings of the Sixth International Congress of Phonetic Sciences, Prague. (Pargue: Academia, 1970).

Pisoni, D. B. and S. B. McNabb. (1974) Dichotic interactions of speech sounds and phonetic feature processing. Brain Lang. 1, 351-362.

Porter, R. J., Jr., and C. I. Berlin. (1975) On interpreting developmental changes in the dichotic right-ear advantage. Brain Lang. 2, 186-200.

Repp, B. H. (1975a) Dichotic forward and backward "masking" between CV syllables. J. Acoust. Soc. Amer. 57, 483-496.

Repp, B. H. (1975b) Dichotic masking of consonants by vowels. J. Acoust. Soc. Amer. 57, 724-735.

Stevens, K. N. and D. H. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. J. Acoust. Soc. Amer. 55, 653-659.

Studdert-Kennedy, M. (1975) Two questions. Brain Lang. 2, 123-130. [Also in Haskins Laboratories Status Report on Speech Research SR-41 (1975), 51-57.]

Studdert-Kennedy, M. and D. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

Studdert-Kennedy, M., D. Shankweiler, and S. Schulman. (1970) Opposed effects of a delayed channel on perception of dichotically and monotically presented CV syllables. J. Acoust. Soc. Amer. 48, 599-602.

Summerfield, A. Q. and M. P. Haggard. (1974) Perceptual processing of multiple cues and contexts: Effects of following vowel upon stop consonant voicing. J. Phonetics 2, 279-295.

Zlatin, M. A. (1974) Voicing contrast: Perceptual and productive voice onset time characteristics of adults. J. Acoust. Soc. Amer. 56, 981-994.

The Magical Number Two and the Natural Categories of Speech and Music[*]

James E. Cutting[+]

ABSTRACT

While upper limits of information processing capture the interests of most experimental psychologists, certain lower limits entice those interested in speech perception. Thus, the magical number for speech is not seven but two, manifested most clearly in the phenomenon of categorical perception. Small deviations from twoness are seen in the perception of stop consonants, whereas considerably larger deviations are seen for vowels. Recently, stop-consonant-like results have been obtained for musical sounds differing in rise time, and identified as pluck and bow. Like the categories for stop consonants, those for pluck and bow appear to be natural and not learned: infants as young as two months discriminate the musical sounds in a manner functionally identical to adults. Mechanisms for the perception of both speech and certain nonspeech sounds appear to be opponent-process feature analyzers not under the conscious control of the perceiver.

Eleanor Rosch (1973) found that the Dani, a nonindustrial and nonliterate community in New Guinea, perceive certain colors and shapes in a manner functionally identical to American college sophomores. Her result is interesting because the Dani have no color terms other than those for light and dark and no terms for angular geometric figures. Her methodology is complex and not relevant to speech research; her discussion centers more on the general area of condition than on perception, but her conclusion is central to my theme: there are salient stimuli in our environment that we perceive as prototypes of natural categories. In other words, our perceptual apparatus is geared to perceive certain stimuli better than others, and it warps a somewhat ill-fitting stimulus to be more like its natural prototype. Moreover, going somewhat beyond Rosch, there are distinct perceptual boundaries between these adjacent categories. The categories and boundaries are "natural" because they remain largely unmodified by learning or by environment. Rosch presented convincing evidence that natural categories exist in vision; here, I hope to demonstrate that they are prevalent in audition and are accompanied by equally "natural" boundaries. I will use

---

[*] To appear in Tutorial Essays in Psychology, ed. by N. S. Sutherland (Hillsdale, N. J.: Lawrence Erlbaum Assoc., in press).

[+] Also Wesleyan University, Middletown, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-42/43 (1975)]

findings of speech research to establish particular patterns of results indicative of "categorical" perception, and then search for them in music as well. Before presenting any data, however, I will discuss a theoretical framework in which to consider categories and boundaries.

A. The Magical Number Two in Speech Sounds

For the last two decades a certain segment of the psychological community has been persecuted by an integer. The persistence with which this number plagues those of us interested in speech perception is far more than a random accident. There is, to quote a famous senator (and perhaps a more famous psychologist), a design behind it, some pattern governing its appearance. Either there really is something profound about this number or else we are all suffering from delusions of persecution. Our number, however, is not seven; it is two.

It is no mere trick that I choose to paraphrase the first paragraph of George Miller's famous paper from Psychological Review (1956). Information processing has certainly burgeoned in the twenty years since his paper appeared, and talk of channel capacities and bits of information has since filled many books and articles. One may worry, then, that those of us interested in this smaller integer are somewhat misguided, if not stunted: perhaps each of us is only two-sevenths of a proper psychologist, or perhaps our student subjects are only two-sevenths as bright as most. This is not the case (we hope). Whereas Miller is concerned with an upper limit of perceptual processing, we are interested in a lower limit. In addition, we are interested in the possible benefits derived from binary systems. In information-theory terms, Miller is a three-bit researcher; we, on the other hand, are not even two-bit but rather one-bit researchers.

Psychologists and others have come very late to one-bit research, especially as it is relevant to language. Millennia before engineers and their computer science stepchildren thought in terms of binary electrical circuits, before physiologists discovered all-or-none neural firings, and before geneticists postulated dominant and recessive genes, Greek and Sanskrit grammarians were discovering the magical number two in distinctive features. These binary systems are fundamental to language: "the dichotomous scale is the pivotal principle of the linguistic structure" (Jakobson, Fant, and Halle, 1951:9). Spoken language, in particular, is a house built on the number two (see also Lane, 1967).

Consider some important binary oppositions in speech, using /ba/, as in bottle, as a reference syllable. Much as a dollar sign denotes that numbers are American money, the slashes here indicate that the letters between them are spoken according to the International Phonetic Alphabet. It is reasonable that /ba/ should be considered a central utterance in a scheme of speech tokens. Unlike many speech sounds, the elements /b/ and /a/, and the syllable itself, are nearly universal to all languages of the world. A related syllable, /pa/, as in pod, is also nearly universal. Together, the two consonants /b/ and /p/ are a voiced-voiceless pair and differ only in the relative timing of the opening of the mouth and the initiation of pulsing in the larynx. For /ba/ the timing is nearly simultaneous in English, whereas for /pa/ there is a slight delay in the onset of voicing, which is preceded by about a twentieth of a second of whisper. This distinction is important, because there is no speech sound, or phoneme, that is intermediate between /b/ and /p/.

Another binary pair is /ba/ and /ma/, which differ in manner of production: /ma/ is nasalized, /ba/ is not, but otherwise they are identical speech sounds. When a child says "I have a cold id by doze," we can appreciate the effect of clogged nasal passages on the neutralization of this phonetic distinction. A third pair is /ba/ and /da/, which differ in place of articulation: /ba/ is labial, produced at the lips, and /da/ is alveolar in English, produced by placing the tongue on the alveolar ridge behind the teeth. Just as there is no speech sound between /ba/ and /pa/, there are none between /ba/ and /ma/ and none between /ba/ and /da/.

Until World War II these distinctions were based on little more than three thousand years of intuition about the nature of speech production. Psychologists, wary if not skeptical of intuition and typically more interested in perception than production, did not become interested in speech until the invention of the sound spectrograph. This device transforms sound into a permanent visual record of time, frequency, and intensity patterns. [See Potter, Kopp, and Green (1947), for elegant and detailed examples of sound spectrograms.] Shortly after the invention of this auditory-to-visual transform came its inverse, a device known as the pattern playback, which transforms a visual display into sound. Through a period of interactive experimentation with these two devices, many of the important acoustic cues were discovered that separate speech sounds from one another (see Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967, for an overview). Schematic spectrograms of the four syllables of particular interest here are shown in Figure 1. Since the three pairs are logically orthogonal, they are displayed as if in three-dimensional space. These examples are exactly like those used for the pattern playback, and would be highly intelligible (if somewhat metallic and "unnatural" sounding) when played through that device.

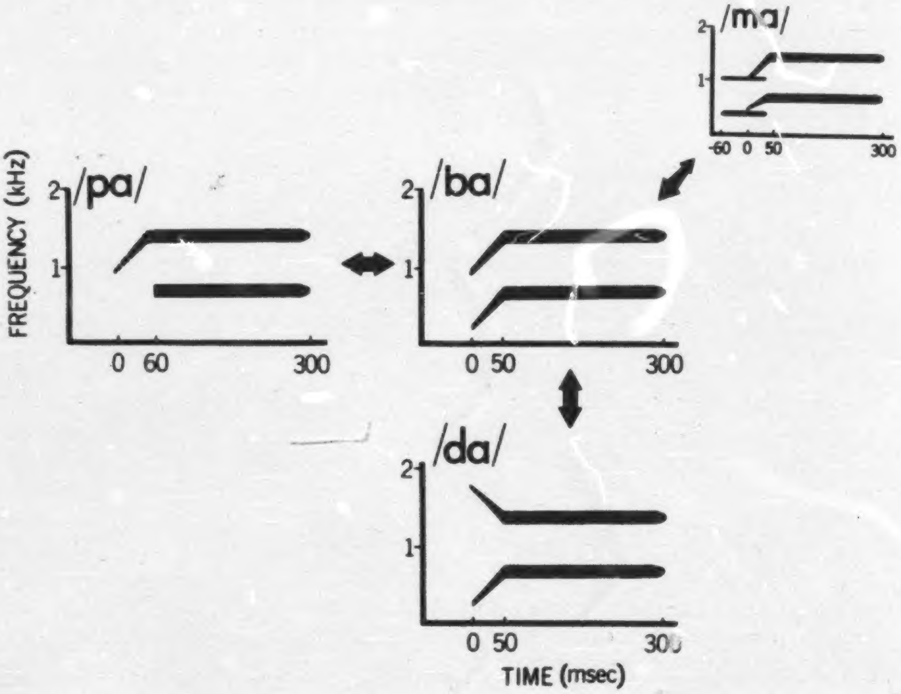

Figure 1: Schematic spectrograms of /ba/ (as in bottle) and three other syllables whose initial consonants differ from /b/ along one phonetic feature.

Observe the acoustic differences between the syllable pairs. Although all pairs are very similar, /ba/ and /pa/, for example, differ in two ways. In /pa/ the first formant, or dark resonance band of lowest frequency, has been cut back from stimulus onset by about 60 msec. Also, the excitation pattern of the second, or higher, formant has changed. Instead of being excited by a periodic glottal source in the larynx, it is excited by aperiodic or noiselike turbulances in the mouth cavity. Natural speech tokens would typically have a third and other higher formants. Whereas the third formant carries some important phonetic information, the fourth and higher formants carry little or none; the first two carry the bulk of the linguistic information load and suffice for these syllables. The syllable pair /ba/ and /ma/ differ mostly in the addition of steady-state nasal resonances to /ma/. They extend from just before to just after the release of constriction at the lips (which creates the formant transitions). For /ma/, however, the first-formant transition is less prominent. The differences between /ba/ and /da/ are perhaps the smallest and conceptually easiest to visualize of the three pairs. In /ba/ the second formant glides upward in frequency at syllable onset, whereas in /da/ the second formant glides downward by about the same amount. It will be instructive to consider this pair in more detail.

Identifying. Humans have little success in producing speech sounds intermediate between /ba/ and /da/. Computer-driven speech synthesizers, on the other hand, can easily be programmed to produce these unlikely sounds. When a seven-item continuum of utterances is generated from /ba/ to /da/, the syllables array themselves as shown in the left panel of Figure 2. When these seven syllables are randomly ordered and presented many times, and when listeners identify each as either /ba/ or /da/, we find our first empirical manifestation of the magical number two. Complementary identification functions show discrete perceptual categories as seen in the upper-left panel of Figure 3. These are actual not idealized data Notice that the first three stimuli in that array are almost always identified as /ba/, and that the last three items are almost always identified as /da/. (Stimulus 4 is perceived as /ba/ about half the time and /da/ the other half.) The stimulus differences appear to be perceived in a discrete rather than continuous manner.

However, one should not be overly impressed with the quantal nature of these complementary functions. Imagine an array of lines tilted at various angles like that shown on the right of Figure 2. If we "read" these lines from left to right, Stimuli 1 through 3 might be considered "ascending" and Stimuli 5 to 7 "descending." Increments of physical difference between members of this visual array are exactly equal in angular degrees, just as increments in the /ba/-to-/da/ auditory series are equal in slope change of the second-formant transition. When the visual stimuli are mounted on cards and viewers are asked to classify each as ascending or descending, we find nicely quantized identification functions shown in the upper-right panel of Figure 3, with only Stimulus 4, the true horizontal, not a member of either category. Clearly, the auditory and visual results are similar, and nothing would appear to be peculiar about speech.

As a further demonstration that identification functions should not be overemphasized, consider what happens when we ask the same listener/viewers to classify the continua into three categories instead of two. The speech-syllable choices here are /ba/, "ambiguous" (not convincing as either stop consonant), .

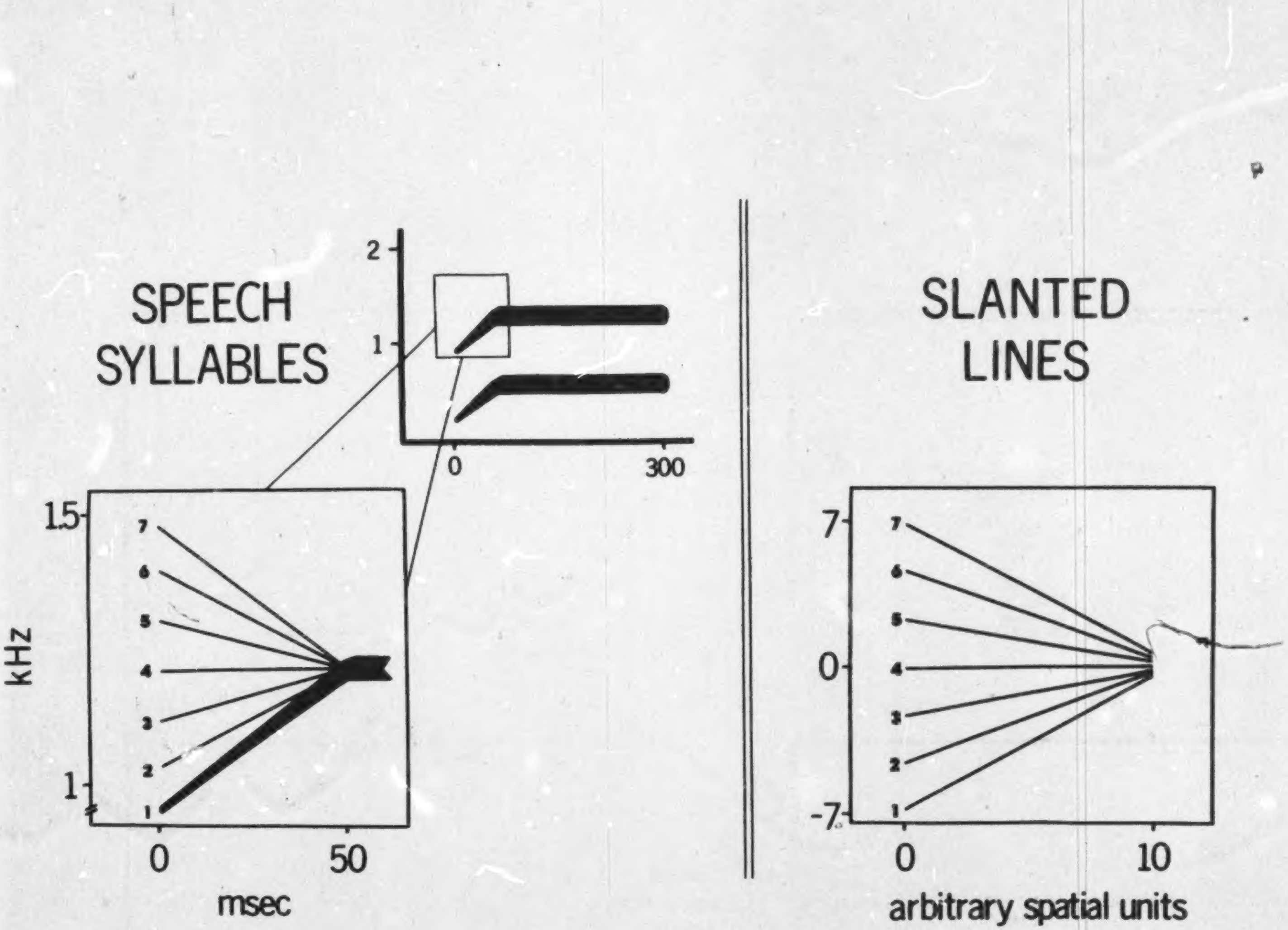

Figure 2: Schematic spectrograms of a /ba/-to-/da/ acoustic continuum, and a display of a companion array of slanted lines.

FIGURE 2

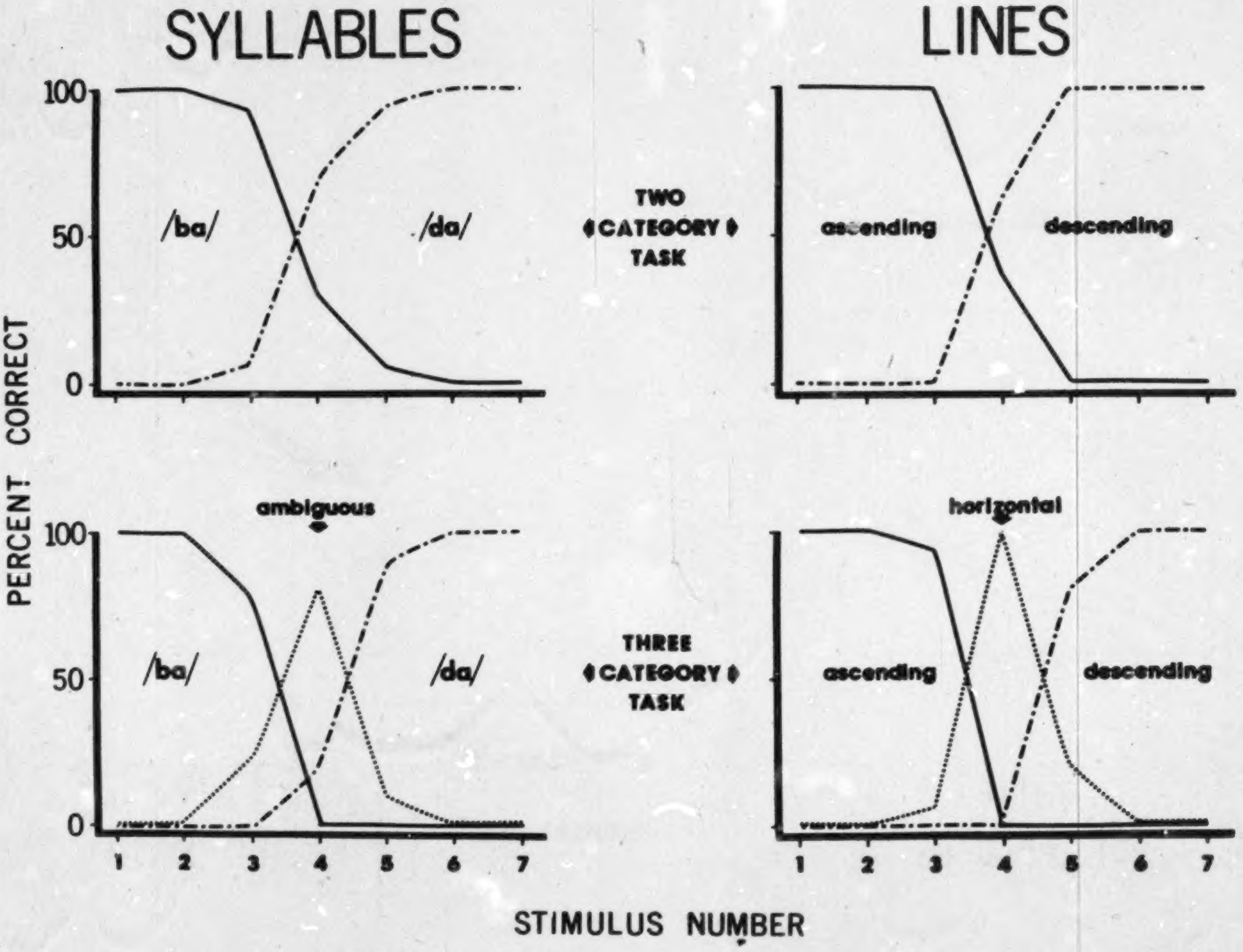

Figure 3: Identification functions for an array of speech syllables and an array of slanted lines (shown in Figure 2) in two conditions: one assigning the items to two categories and the other assigning them to three categories.

FIGURE 3

and /da/; and the slanted-line categories are ascending, horizontal, and descending. Results are shown in the lower panels of Figure 3. Both classes of stimuli yield similar identification patterns, with the third categories supplanting the old boundaries in the two-category tasks. From these results, speech perception would appear to be no different from the perception of objects and events in other modalities. Moreover, the magical number two would seem irrelevant.

Two statements must be made before entertaining the notion that these conclusions are legitimate. First, the two stimulus series in question were judiciously selected. Few acoustic continua generated by a speech synthesizer appear to have phoneme boundaries with near-zero slope in the second-formant transition: /ba/-to-/da/ is closer to being an exception than the rule. Deviations from the peculiar regularity in this syllable array, such as that found in a /bi/-to-/di/ acoustic continuum (as in beam to deem), would be much more difficult to model in a visual continuum. Second, we should consider the nature of the middle categories in each set of responses. Intuitively they seem quite different. The middle visual category would appear psychologically more real than its neighbors. Indeed, the terms ascending and descending are derived with reference to horizontal. The middle speech syllable category, on the other hand, is a tenuous if not bogus domain. Certainly, /ba/ and /da/ are not derived perceptually with reference to an ambiguous stimulus that is difficult if not impossible to pronounce. In short, we see horizontal lines every day; we do not "hear" ambiguous speech sounds. Just as with a Necker cube, the percept flips one way or the other: it is either /ba/ or /da/, and rarely anything else, unless one asks the subjects to perform the unusual task of "ambiguating" the syllables as I have done.

Discriminating. Given these clues that a /ba/-to-/da/ acoustic continuum is perceived somehow in a unique and quantal manner, we should look to a second and more important manifestation of the magical number two--nonlinearities in discriminability.

If a listener/viewer is asked to compare two members of one of the arrays of stimuli used thus far, how accurate are her responses? For purposes of uniformity, both arrays of stimuli are presented in a sequential discrimination task: the first stimulus is presented, followed by a silent or blank interval of one second, followed by the second stimulus (either identical to the first or two steps removed along the physical continuum). In this manner, along with item-pairs that are identical, Stimuli 1 and 3, 2 and 4, 3 and 5, 4 and 6, and 5 and 7 are compared. Subjects are asked to report whether the two items are the same or different. Only the "different"-pair results are of interest here and are shown in Figure 4; few errors occur on "same"-pair discriminations in this type of task. Notice the sharp discrepancy between the two darker functions. The speech-syllable data, shown in the top panel, demonstrate a sharp peak in discriminability at the Stimulus-3/Stimulus-5 comparison that rapidly tapers to lower-than-chance performance at either end of the continuum. The slanted-line function, on the other hand, is at or near 100 percent performance throughout the stimulus range.

Comparing these discrimination results with the two-category identification functions superimposed on them, we see that for the speech items there is a correspondence between the crossover of the complementary identification

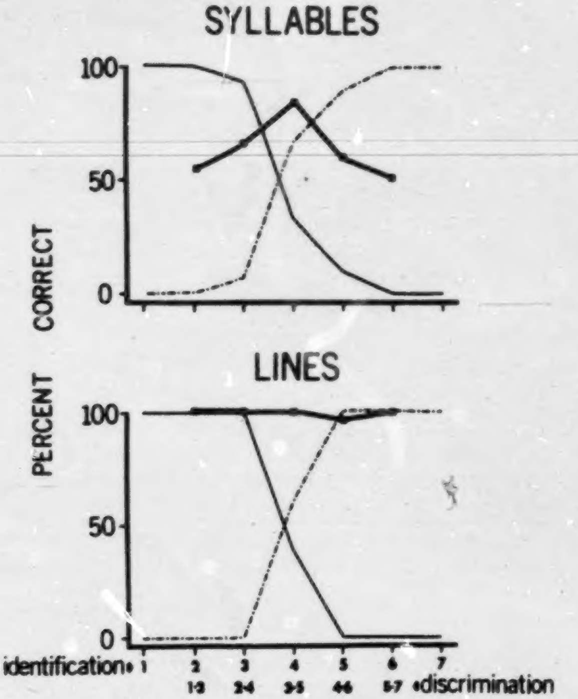

Figure 4: Two-step discrimination functions for speech syllables and slanted lines, superimposed on their respective identification functions.

curves and the peak in the discrimination function. Labelability changes inversely with discriminability. Items can only be perceived as distinct from one another when they have different names. This nonlinearity lies at the heart of the interest in the magical number two, it is called categorical perception, and it is in sharp contrast to the performance on the same task for the slanted-line stimuli. Acoustic differences between speech Stimuli 1 and 3 or 5 and 7, for example, are typically inaccessible to the listener. However, the magical number two is not as absolute as it seems here, with two discrete categories and a perceptually distinct boundary between them. It is necessary to consider certain systematic deviations from strict twoness.

B. Plus or Minus Two Fudge Factors

Just as Miller (1956) has his small margin for error, a fudge factor of plus or minus two, we speech researchers also have ours. It is considerably smaller; it is difficult to scale down to size in terms of numerical deviations from our magical integer; in fact, it may be difficult to determine whether it is actually positive or negative. What is clear is that it manifests itself, roughly, in two "sizes," one trifling and the other not-so-trifling.

The trifling deviation from the magical number two, and the strict categorical perception it implies, can be seen using the same /ba/-to-/da/ continuum. When listeners are asked to attend to Stimuli 1 and 3, for example, both of which are identified as BAH nearly 100 percent of the time, and they are asked to judge which of the two is more BAH-like--or in Rosch's (1973) terms, which is the prototype--they will typically determine that Stimulus 1 fits the bill. They will do this, however, only after they have laughed at the experimenter for asking them to perform such a ridiculous chore, only after she has reassured them that there really is a difference between the items, and only after she has cajoled and exhorted them to do the best they can. Even after these machinations, their performance is rarely close to being perfect. This special tutoring of within-phoneme-class acoustic differences does not appear to transfer to other stimuli, and often is not useful as pretraining for other tasks with the same stimuli. The fact that subjects can report differences between two different tokens of /ba/ is more a testament to what the human perceptual apparatus can do in an unusual situation rather than what it does do in a normal situation. In all fairness, however, it should be noted that this deviation from the magical number two is more trifling in size than in theory. Its discovery by Barclay (1972) was a blow to some stricter views of speech perception.

The not-so-trifling deviation from the magical number two requires another set of stimuli, and it is theoretically even more important than the first fudge factor. Shown at the top of Figure 5 are the endpoints of an acoustic continuum of vowels from /i/ as in heat to /ɪ/ as in hit. Between them, one can easily generate five intermediate stimuli, thus creating a seven-item array with equal increments of acoustic change between all members. Here, instead of changing slopes of transitions, the frequencies of entire resonances are changed, increasing in value from /i/ to /ɪ/ for the first formant and correspondingly decreasing in value for the second and third formants. (The addition of the third formant here increases intelligibility, but the array would yield nearly identical results without it.) When these items are randomly ordered and presented many times to listeners, results show quantal identification functions similar to those shown at the top of Figure 3 for consonants and for slanted lines. That is, Stimuli 1 through 3 are identified as /i/, Stimuli 5 through 7 are identified as /ɪ/, and only Stimulus 4 is ambiguous between the two. Discrimination results, however, reveal a pattern unlike those for consonants or for slanted lines. They are shown in the lower panel of Figure 5.

Notice that the vowel discriminations lie intermediate between the previously discussed consonant and slanted-line functions. There is a "peak" in the function at the Stimulus-3/Stimulus-5 comparison but the "troughs," or regions of poor discriminability, are not nearly as "deep" (close to zero percent performance) as those for stop consonants. The reason for this appears to be that within-category differences between two tokens of the same vowel remain in short-term memory long enough for accurate comparisons to be made. Performance on these comparisons, however, is still worse than for those made at the phoneme boundary. The reason for the trough-peak difference is related to the way in

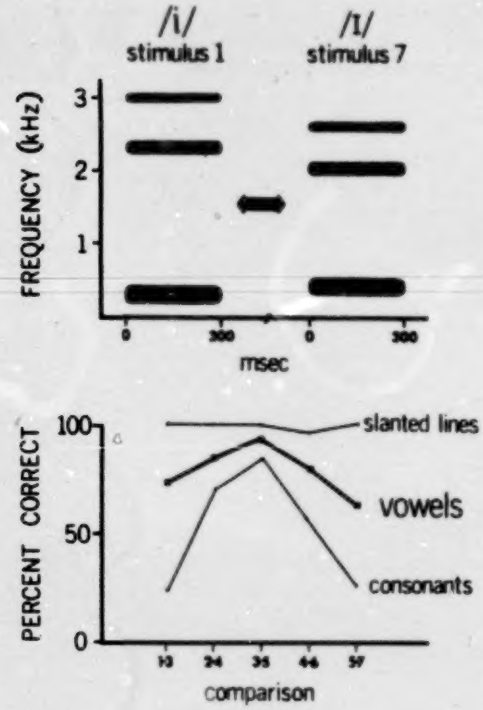

Figure 5: Schematic spectrograms of /i/ and /ɪ/ (as in eat and it), endpoints of a seven-item acoustic array, and the discrimination function for that array compared against the slanted lines and stop consonants.

which the information is encoded. For speech stimuli the height of the peaks in any discrimination function can be taken as a measure of the strength of phonetically coded information, and the depth of the troughs in relation to those peaks can be interpreted as the relative strength of acoustically coded information. Only the phonetic code is relevant to short-term memory as it is usually defined; the acoustic code fades much more rapidly. Differences between the consonant and vowel functions at within-phoneme-category comparisons are one assessment of the magnitude of this not-so-trifling deviation from the number two. Some raw acoustic information about vowels is available for comparison purposes; practically none is available about the consonants.

One feature of these discrimination results that I have not discussed so far is the effect of the duration of the silent interval between the two stimuli in the sequential discrimination task. For the vowel stimuli, this interval is vital for determining the depths of the troughs in the functions. If the interval is shortened from one second to a quarter of a second, listener performance on within-category comparisons of Stimuli 1 and 3 or of Stimuli 5 and 7 will increase to as much as 85 or 90 percent. If, on the other hand, this interval is lengthened to as much as two or three seconds, listener performance on these same comparisons will decrease to as low as 40 to 50 percent. There is no such effect of silent or blank interval on consonants, and probably none for slanted lines (although I have not done the experiment). For consonants, in particular, the within-category acoustic information appears to be lost prior to the onset of the second stimulus in a to-be-discriminated pair, regardless of how short that interval may be. Perception of consonants, then, is almost instantaneously phonetic. Practically no raw acoustic husk remains in memory. In Rosch's (1973) terms, our perceptual apparatus warps all stop-consonant stimuli that fall within a phoneme category, changing them into a perceptual prototype. Once that prototype is internally registered, the nonprototypic vagaries of the stimulus are largely inaccessible to consciousness except through experimental exhortations like those discussed earlier. A more careful account of these phenomena and the extent to which they occur has been developed by Pisoni (1973, 1975; Pisoni and Lazarus, 1974; Pisoni and Tash, 1974), and the interested reader should refer to those articles.

There are many psychological differences between stop consonants and vowels. I feel that it is incorrect, however, to say one class of phonemes is more "speechlike" than the other. Nevertheless, since the perception of stop consonants adheres more to the magical number two, it is those stimuli that we will consider in depth in this chapter. Before continuing with the stops, however, we must consider the possibility of categories and boundaries elsewhere in audition.

C. The Magical Number Two in Nonspeech Sounds

For years, those of us interested in speech and the number two (implied by the phenomenon of categorical perception) have called those auditory events that are not speech "nonspeech." Underlying the use of this handy and linguo-centrically biased term is the belief that speech is somehow different from all other auditory events, just as it is different from the perception of slanted lines. Most of us, I think, still believe this to some degree. I do. But, whereas we used to be armed with an arsenal of empirical data supporting the uniqueness of speech processing, lately we have begun to strip ourselves of these findings. The most important weapon in our arsenal and seemingly the most invulnerable to attack was categorical perception.

Shortly after the initial formulation of categorical perception--that it involves a discontinuity in discrimination functions for stimuli equally spaced along a physical continuum--arose the issue of "acquired distinctiveness." Expressed as a question in its simplest form: Are the discrimination peaks learned for stop-consonant stimuli? Do children, for example, acquire the distinct speech categories, or are they innate? A dozen years after the question first arose, it was answered conclusively, and that answer is discussed in Section D. At the time, however, it was not possible to test young infants, so the question was asked in another form: Can nonspeech discrimination peaks be acquired through training? Perhaps the process of acquiring categories and boundaries follows a developmental trend: initially, all stimulus pairs may be equally discriminable to the untrained listener and her discrimination function would be "flat" and moderately above chance level throughout the range of the continuum; only later, with training, would a peak appear in this function, and perhaps the troughs would correspondingly drop within each category.

Harlan Lane, an early proponent of this view, found that certain subjects listening to certain complex nonspeech sounds (spectrographically inverted speech patterns) acquired discrimination peaks through a simple training procedure (Lane, 1965). Pisoni (1971), in a careful replication with similar stimuli, found this to be true for a few selected subjects, but generally not true. Moreover, Studdert-Kennedy, Liberman, Harris, and Cooper (1970) found even

Lane's selected data unconvincing: whereas there were peaks in his functions, Lane found few deep troughs and less correspondence between discrimination and identification functions than would be desired. As seen when comparing the perception of vowels and consonants in Figure 5, troughs are vital. If the distinctive nature of the peaks can be acquired through training, and it is not entirely clear that they can be, the troughs do not appear to be learned: there seems to be no pneumatic trade-off between the acquisition of peaks in training and the loss of ability to discriminate within a category.

But the Lane and the Pisoni stimuli were not "natural" nonspeech sounds; nor are sine wave tones and other more familiar psychoacoustic stimuli "natural" in any real sense. Are there commonly occurring stimuli in our environment that are perceived categorically and that obey the laws of the magical number two? An obvious candidate here is musical sounds. They are natural at least to the extent that they rely on simple mechanical action of easily fashioned materials. Locke and Kellar (1973) varied the middle component of triadic chords in search of categorical perception, and found some categorical tendencies in musically trained listeners, but few in musically naive listeners. Their results were promising, but with two important drawbacks. First, the discrimination functions for the musically trained listeners were more similar to the vowel function shown in Figure 5 than the typical stop-consonant function beneath it. Second, and more damning, is the fact that extensive musical training seemed to be a requisite for even these vowellike functions. Again, we are back to "acquired distinctiveness," and to the lack of sufficient troughs in the discrimination functions.

More recently, categorical perception has been found in a musically relevant dimension and the results meet all the requirements for binary processing according to the laws of the magical number two (Cutting and Rosner, 1974; Cutting, Rosner, and Foard, 1975). The dimension is that of attack, or rise time. Rapidly rising sawtooth waves, for example, sound like the plucking of a stringed instrument, such as a guitar; more slowly rising sawtooth items sound like the bowing of a similar instrument, such as a violin. Oscillograms of token "pluck" and "bow" sounds are shown in Figure 6. Rise time cannot be systematically varied when playing actual musical instruments, but it can be varied readily on a Moog synthesizer. We chose to vary the rise time from 0 to 80 msec in 100-msec increments for several continua of musical sounds. Note that such variation is minor in magnitude compared to the long and tapering offset of the stimuli.

When these stimuli are placed in the same paradigms mentioned previously for the speech syllables /ba/ and /da/, they yield remarkably similar results. Sawtooth wave items with less than 40-msec rise time are identified as pluck nearly 100 percent of the time, stimuli with rise times greater than 40 msec are identified as bow nearly 100 percent of the time, and only the 40-msec rise time stimulus is ambiguous (identified as pluck about 40 percent of the time and as bow 60 percent of the time). When these stimuli are placed in the sequential discrimination task, items within a category sound identical to the listener: asked to judge whether pairs of stimuli are the same or different, listeners report that items of 10- and 30-msec rise time and of 50- and 70-msec rise time are the same item more than 75 percent of the time, well below chance. Only when items with 30- and 50-msec rise times are compared do listeners perform well, and here they make only about 15 percent errors. These results clearly

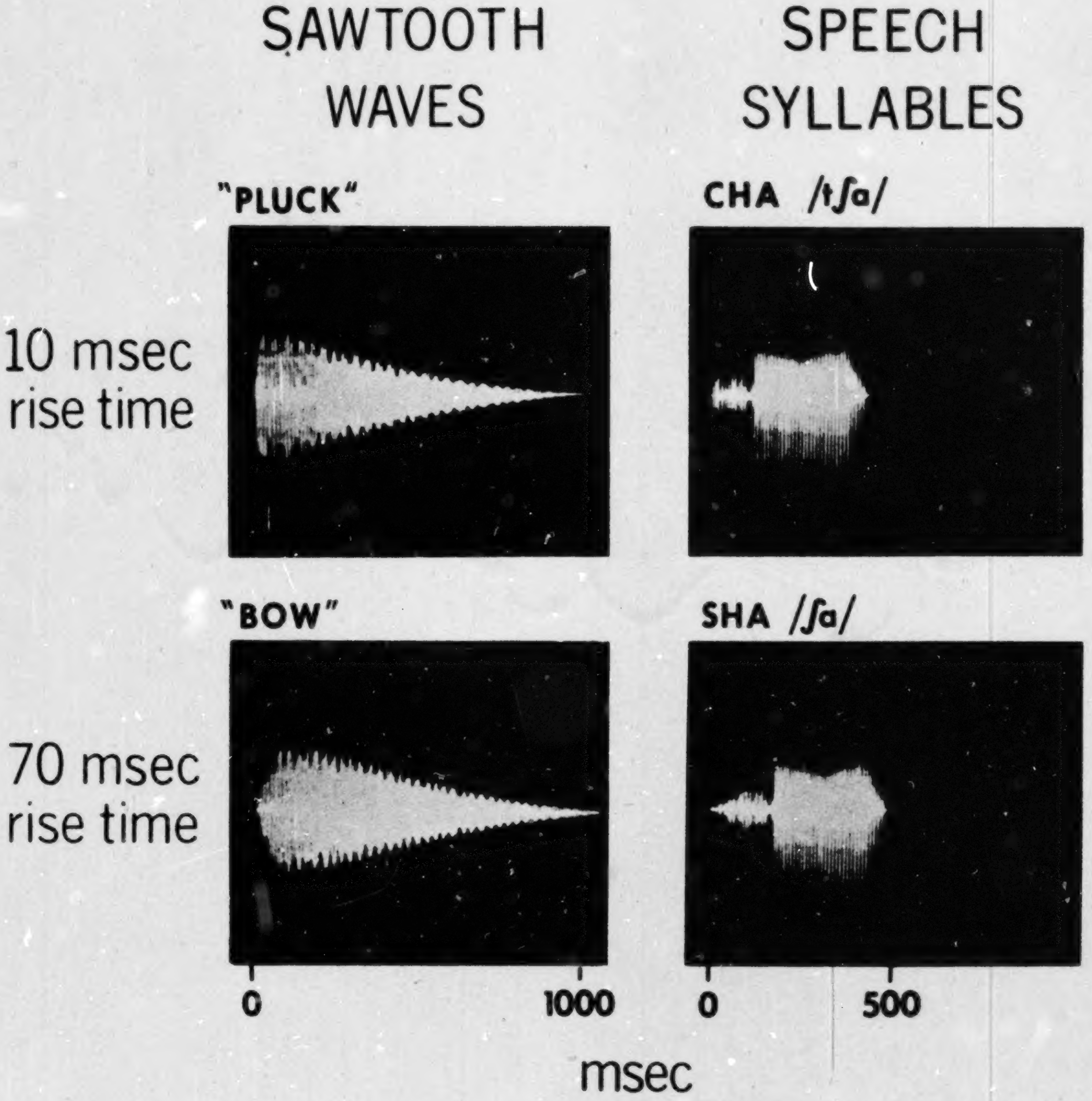

Figure 6: Oscillograms of token pluck and bow stimuli, compared against token /ʃa/ and /tʃa/ items. Both pairs of stimuli differ in rise time.

FIGURE 6

indicate that there are quantal identification functions for pluck and bow sounds, that there is a peak in the discrimination function lying astride the crossover point of the complementary identification functions, and that the discrimination function falls off into deep troughs at either side of the peak. Thus, the magical number two reigns in nonlinguistic domains as well as in speech.

Six other aspects of our data are important concerning categorical perception of these musical sounds. First, the time interval between the items in a discrimination pair matters not at all: performance on within-category pairs, for example, is no better when the two stimuii are separated by 250 msec than when they are separated by nearly two seconds. These results strongly suggest that the musical sounds adhere to the magical number two more strictly than do vowels and that the "fudge factor" is likely to be of the trifling rather than the not-so-trifling size.

Second, the results do not appear to be mediated directly by the labels pluck and bow. In our initial study we were careful to administer the test in two conditions. In one we carefully tutored subjects in the use of the terms pluck and bow, playing extreme tokens from the continuum several times before they participated in the identification test. The identification test gave them an additional 15 minutes of practice using the terms before they listened to their first discrimination pair. In a second condition the subjects started right away with the discrimination test, did not hear practice items, and were not told of the labels pluck and bow. The results for the two groups were essentially identical and suggest that the labels pluck and bow have little to do with the perceptual process.

Third, one might think that our results are robust because the stimuli emulate common musical sounds or because the stimuli have complex spectra. Would such results occur for simpler auditory events, varied in the same manner, that do not sound like convincing tokens of musical sounds? We varied rise times of sine wave stimuli and found the same pattern of identification and discrimination results as for the sawtooth items. These results indicate that the perceptual process involved here is more fundamental than just a music-processing system might be. Sine waves with rapid rise times sound vaguely like a flute played staccato style; but sine waves with less rapid rise times are not at all convincing as notes from a flute played in a more legato style.

Fourth, one may be suspicious of our stimuli, since we played them over loudspeakers and earphones. Because certain members of the musical arrays have very rapid rise times, they are likely to induce clicks into the transduced signal; that is, the response characteristics of the broadcasting devices may be sluggish enough to produce audible short-bursts at the beginning of the items. One fear is that the presence or absence of such clicks might correlate perfectly with the perceptual categories: pluck equals presence of a click, bow equals no click. To ensure that this artifact did not account for our results, we inspected our stimuli after they were played through a loudspeaker, redisplaying them with high resolution on a computer-controlled oscilloscope. We found such small perturbations in the signals only for the 0- and 10-msec rise time stimuli. However, items with 20- and 30-msec rise time, which were members of the same perceptual category, did not have these irregularities.

Fifth, we found that the long tails on the pluck and bow stimuli are necessary for the successful identification and discrimination of the items. When the items are trimmed to 250 msec in duration by simply lopping off the last 750 to 850 msec of the items, identifiability and discriminability are markedly impaired. At first, this seemed rather bizarre to us: the final three-quarters of the stimulus does not appear to carry any information about stimulus onset and would seem unnecessary for maintaining musical integrity of the sounds. This view is clearly incorrect. Moreover, we should not have been surprised at these results. When speech items such as /ba/ and /da/ are severely trimmed from 300 msec to 40 or 50 msec, removing only the steady-state vowel, they often cease to sound like speech stimuli and are unlabelable by many listeners. The integrity of the syllables is thus violated, and in a manner similar to the truncation of our pluck and bow items.

Sixth, and perhaps most interesting of all, is the fact that rise time is not only a cue for the distinction of musical items, but it is also used as a cue in speech: CHA (/tʃa/), as in chop, has a very rapid rise time in its fricated (or noiselike) portion, whereas SHA (/ʃa/), as in shop, has a much more gradual onset. Tokens of these speech syllables are shown in Figure 6 next to the pluck and bow items. When an array of these syllables is generated on a computer-driven speech synthesizer, and when they are inserted into the same paradigms as we have discussed thus far, listeners yield patterns of identification and discrimination results that are nearly identical to those for the pluck and bow items. We find it compelling that a single cue, rise time, is used to distinguish categories inside and outside of speech. While speech production and speech perception are unique to man, we should not expect all speech processing mechanisms to be unique as well. In evolutionary terms, it would have made sense to build a speech processing system on underlying and already existing auditory faculties. It seems reasonable that at least some of the binary distinctions on which speech is built would be based on binary auditory distinctions. We suggest rise time is one of them (Cutting and Rosner, 1974).

To account for discontinuities in the discrimination functions of stimulus arrays /ba/-to-/da/ and /i/-to-/ɪ/, many speech researchers have thought in terms of phonetic and auditory memories. The peaks in these functions have been taken as a measure of the strength of phonetic memory similar to the more commonly known short-term memory; and the troughs, in relation to the peaks, are taken as a measure of an auditory memory similar to what is often called echoic memory. The categorical perceptions of pluck and bow stimuli jar this view somewhat. The notion of an auditory memory accounting for the troughs remains unchallenged. However, the notion that a phonetic memory underlies the peaks in a discrimination function must be cast aside. The peak in the pluck and bow discrimination function can in no way be thought of as phonetic. Instead, this higher-level memory may be reserved for highly coded decisions about auditory signals; pluck versus bow or /ba/ versus /da/ would both qualify here. It is relatively easy to understand why speech sounds are categorical and coded into a phonetic string rather than left as raw acoustic information. The memory storage capacity required for one second of high-quality speech (such as reproduced on a tape recorder) would be 40,000 bits of information, whereas the storage capacity required for one second of phonetically coded speech would be only about 40 bits of information, plus the necessary subroutines to decode that string (Liberman, Mattingly, and Turvey, 1972). Clearly, in terms of a thousand-to-one savings in storage capacity, it makes sense to code speech phonetically and at a rather rapid rate so that it can be comprehended. The rub, however, is to understand why the same system appears to code and categorize musiclike sounds when such a savings may not be needed. The answer is necessarily indirect, and must first take us back to the notion of "natural" categories and their apparent function.

D. Naturalness of the Magical Number Two in Speech and Music

Thus far I have presented what should be compelling evidence of discrete categories in speech and music, but I have said nothing of their "naturalness." Rosch concluded that certain color and shape categories are natural because they appear to remain largely unmodified by the presence or absence of language terms for them. To find exactly parallel results for speech items is difficult. Speech syllables have the unique property of providing their own nonarbitrary labels: /ba/ is BAH and /da/ is DAH, and they are pronounced and labeled as such by (almost) all peoples of the world. We are, therefore, forced to use a different technique for assessing naturalness of speech categories and boundaries, and we must use the same method for those in music.

The Oxford English Dictionary defines "natural" as: "present by nature; innate; not acquired or assumed." If speech categories are natural by this definition, all humans should be born with the ability to use them. One approach for determining whether or not the perception of these categories is innate is to test young infants. For reasons of practicality, the infants tested have been from one to four months old. The assumption here is that these infants will have had little if any opportunity to learn much about their to-be-native language and that any results they yield are characteristic of those capabilities that are genetically "wired-in."

Speech categories. It should be clear that one cannot ask infants to identify /ba/ and /da/. Young children typically cannot produce such differences in a systematic and controlled fashion until they are many months older. Therefore, it is out of the question to try to obtain identification functions. One can determine, however, whether or not infants can discriminate speech sounds and discriminate them in a manner approximating that for college-aged subjects. This is exactly the approach of Peter Eimas and his colleagues (see, for example, Eimas, Siqueland, Jusczyk, and Vigorito, 1971; Eimas, 1974; Cutting and Eimas, 1975) in a series of pioneering studies.

It is one thing to ask an infant to discriminate two speech sounds, but it is another to pose that question in a manner for which he can give a suitable and measurable response. Eimas has used a conditioned nonnutritive sucking procedure; others have used heart rate (Morse, 1972). In the Eimas and Siqueland procedure the infant is given a hand-held nipple on which to suck. Instead of transducing nutrients, it transduces pressure to a pressure-sensitive apparatus, which in turn triggers, for example, the speech sound /ba/. The more frequent the high suction responses of the infants, the louder (or, in another procedure, the more frequent) the speech sound is presented against background noise. The infant quickly learns this association and is quite willing to make several hundred sucking responses over the course of about ten minutes merely to hear the same sound repeated. The time-frequency course of the infant's responses is of particular interest. Over the course of about three minutes after the initial learning of the association, the infant increases his responses to a peak of as much as 50 or 60 per minute, well above a preassociation nonnutritive baseline. Shortly thereafter the infant seems to tire of the situation and responses taper off rather dramatically in the following two minutes.

After a drop in responses of at least 30 percent, one of three things happens to the infant. In one control condition the infant continues to hear the same stimulus, Stimulus 3 in Figure 2, say, over and over again. Responses here continue to approach asymptote at or below the baseline rate. In the experimental condition, however, the stimulus is shifted to /da/ (Stimulus 5) and the infant's responses begin to increase again, only beginning to fall the third or fourth minute after the stimulus change. Most important is the second control condition. Here the stimulus shifts from Stimulus 3 to Stimulus 1, but both are identified as /ba/ by adults. In other words, this change is physically just as great as the across-boundary shift, but both stimuli lie within the same category. As in the first condition, the infant's responses continue to approach asymptote. All these trends are shown in Figure 7.

Three aspects of these infant data are interesting when compared with the adult data discussed in previous sections. First, the across-category stimulus shift in the experimental condition here corresponds to the peak in the discrimination functions seen in the top panel of Figure 4. In the infant's case, the dishabituation of the sucking response is taken as evidence that he perceives that a new stimulus has been presented, one that deserves more attention and subsequently more sucking responses. Hence, like adults, infants as young as one month can perceive phonetically relevant features. Second, the within-category stimulus shift in the crucial, second control condition corresponds to the troughs in the adult discrimination function. Continued habituation of the sucking response is taken as evidence that the infant did not perceive that a "new" stimulus had been presented. Just as for the adult, the infant may have merely regarded the second stimulus as identical to the preshift stimulus. Thus, like adults, infants as young as one month cannot perceive phonetically irrelevant changes in acoustic features even when they are identical in magnitude to the across-category, phonetically relevant change. Third, and more speculative than the previous two points, is that the difference between the functions for the no-shift and within-category-shift conditions suggests that even for very young infants there is a trifling-sized fudge factor that modifies the magical number two. Although the difference between these two groups has never been significant in a single study, the trend is unmistakable: the infants in the within-category-shift condition attenuate their habituation rate slightly more than the no-shift group.

Do infants perceive categorically the same speech continua that adults perceive? It appears that they probably do, and maybe even more so. For example, they yield results functionally identical to those in Figure 7 when discriminating a voice-onset-time continuum from /ba/ to /pa/, and when discriminating /ra/ from /la/ cued only by changes in the third-formant transition. This second result is important since the difference is one that native Japanese-speaking adults cannot perceive and do not have in their language. Such a phenomenon presents us with the tantalizing notion that infants may be superior to adults in perceiving certain speech-relevant dimensions. It suggests that, while it is true that the distinctiveness of speech categories is not acquired by a learning process, it may also be true that certain potential distinctions are lost when unused by the developing child. One might consider this process "acquired indistinctiveness." It might even appear to support some of Lane's (1965)

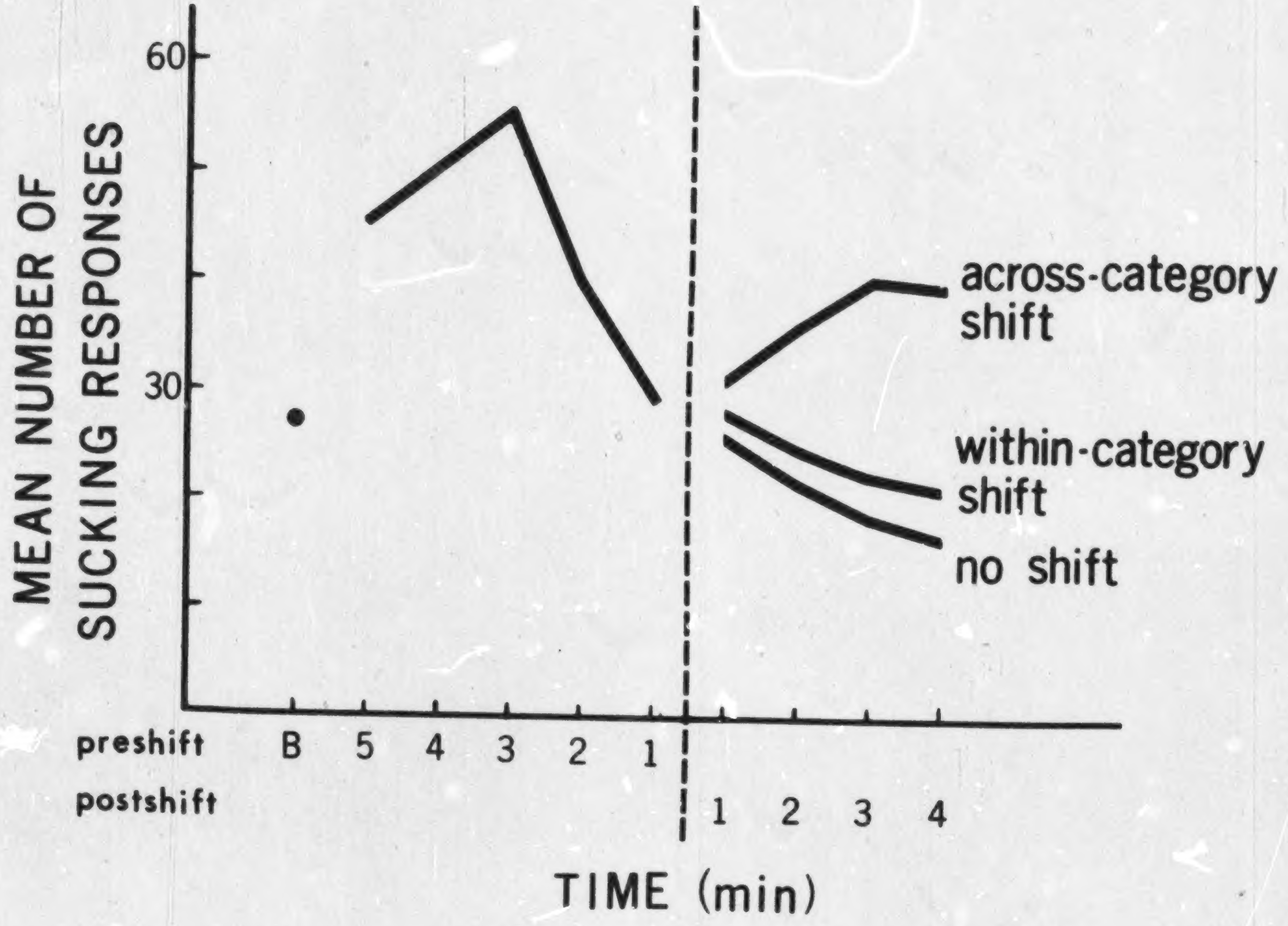

Figure 7: Schematic data display of the time course of infant sucking responses to auditory stimuli in the Eimas paradigm.

FIGURE 7

original contentions. It does not, however, since it is not the troughs of the discrimination functions that get deeper (indistinctive) but rather the peaks themselves that disappear.

Music categories. Are pluck and bow sounds discriminated categorically by infants? They are, as we have recently discovered (Jusczyk, Rosner, Cutting, Foard, and Smith, 1975). The stimuli used in this study were selected from those used in the adult studies (0-, 30-, and 60-msec rise time items) plus an additional stimulus from the original set (with 90-msec rise time). Each of 18 two-month-old infants was run in two conditions: all participated in a condition involving the cross-boundary, 30- and 60-msec items, and six infants each were involved in three control conditions. One group performed the no-shift control; that is, each infant continued to listen to the same stimulus throughout the experimental session. A second group listened to the 0- and 30-msec rise time items, and the third group listened to the 60- and 90-msec rise time items. Counterbalancing of pre- and postshift stimuli was observed, as well as counterbalancing the order of the experimental (30 to 60 msec) and control conditions (no shift, 0 to 30 msec, or 60 to 90 msec).

Results were compelling. Seventeen of the 18 infants demonstrated a higher sucking response rate in the cross-category-shift condition (30 to 60 msec) than in the control condition, regardless of which of the three groups they belonged to. The general patterns of responses plotted over time was functionally identical to those shown in Figure 7: habituation of the response continued in the no-shift condition and in both within-category-shift conditions, while dishabituation occurred for the across-category-shift condition.

The adult boundary for these sawtooth wave pluck and bow stimuli is at about 35- or 40-msec rise time; it is clear that the infant boundary is near this same mark. It should also be clear that both speech categories and "nonspeech" (musiclike) categories and boundaries are innate to humans: one can think of little evidence supporting the notion that these young infants could have acquired the distinctive categories demonstrated here. Thus, these categories and boundaries appear to be "natural" according to the strictest possible psychological interpretation of that word.

An interesting question now arises. These categories and boundaries seem innate to humans, but are they innate to other animals (e.g., primates) as well? In other words, is there any evolutionary continuity in the development of the mechanisms behind these effects? Morse and Snowdon (1975), for example, have tested rhesus monkeys in the discrimination of speech syllables such as those discussed here. Their results do not support a strong form of categorical perception in infrahuman primates, but there are some categorical tendencies. Would rhesus monkeys discriminate pluck and bow sounds in a manner as striking as adult and infant humans? Alas, we don't know yet. Results here will be interesting regardless of the outcome. If rhesus monkeys (or higher nonhuman primates) do discriminate these musiclike sounds, we will have evidence that the perception of rise time, which is used to cue certain speech categories, has an evolutionary history older than that of speech. This would support the view that certain categories in speech were built on preexisting nonlinguistic categories (Cutting and Rosner, 1974). If the nonhuman primates do not discriminate pluck and bow sounds, we will have evidence that the perception of rise time, and the categories and boundaries in speech and music that it cues, evolved relatively late. Moreover, waxing slightly toward the philosophical, music perception (and subsequent music appreciability) has a recent evolution of its own.

In the first two-thirds of this paper I have presented considerable evidence to demonstrate the existence, and the "naturalness," of certain perceptual categories and boundaries in speech and music. I have not attempted to be overwhelming in breadth: few stimulus dimensions in speech were discussed and only one in music, partly because of the lack of knowledge in this relatively new field. In the rest of the paper I will consider the nature of the mechanisms behind these manifestations of the magical number two.

E. Mechanisms behind the Magical Number Two

Very recently a new paradigm has emerged in the field of speech perception. The technique is known as selective adaptation. Its roots are in vision research and in the mapping of the architecture of brain-cell function during perception. However, it is logically similar in many ways to the much older and better-known phenomena associated with visual afterimages. I will use those photochemical data as a basic framework for discussing the perceptual effect in speech.

It is commonly known that if a person stares at a patch of blue color for about 15 to 30 seconds, and then stares at an illuminated white wall, she will see a patch of yellow color with the same contour as the original blue patch. This effect is known as a chromatic afterimage. It is best explainable in terms of opponent-process mechanisms, as first championed by Hering. Blue is thus viewed as the opposite color to yellow, at least at some stage of photochromatic analysis subsequent to excitation of the cones on the retina. Blue and yellow appear to synapse, if you will, onto the same cell bodies, but one color in an excitatory fashion and the other in an inhibitory fashion. Staring at a blue patch for many seconds fatigues the visual system such that when given a neutral stimulus (white), the viewer will perceive for a brief period of time that stimulus as being of the opposite color (in this case yellow). In a reciprocal fashion, staring at a yellow patch will yield a sensation of blue when the viewer is presented with a postadaptation neutral field.

One cannot stare at a speech syllable. Thus, the experimental adaptation situation must be changed to a degree for such auditory sounds. Since the auditory signal fades rapidly [which Hockett (1960), for one, views as a blessing], it must be presented over and over, perhaps as many as 100 or 200 times, to continually refresh the perceptual "image." If this adapting stimulus is /da/, for example, and one is presented with a neutral stimulus near the /ba/-/da/ boundary (Stimulus 4 shown in Figure 2), the listener will perceive that neutral stimulus as being a good exemplar of /ba/. One may view the stimuli /ba/ and /da/ as being "opposites" of one another along the dimension of change in the second-formant transition. Hence, just as yellow is the opposite of blue, /ba/ is the opposite of /da/, and after adaptation, a neutral stimulus between the two prototypes will be perceived as being a member of the opposite class. Moreover, after adaptation the physical domain of the category of the adapting stimulus shrinks, while the physical domain of the unadapted stimulus category expands to fill the void.

The adaptation paradigm as used in speech perception is an exhaustive one. Typically, the listener is presented with more than 100 tokens of one of the endpoint stimuli in a speech continuum (in this case the Stimulus 1, /ba/, or the Stimulus 7, /da/) and then given a brief identification test of the array of stimuli. This postadaptation test may consist only of one token of each of the seven stimuli in the array presented in random order. The subject identifies each of these items as /ba/ or /da/. Then another long series of adaptation presentations begins using the same stimulus as before. After this sequence, a second seven-item series of stimuli is presented for the listener to identify. This cadence may be repeated a dozen times or more before the experimental session in completed.

Typical results are shown in the top panel of Figure 8. Given an adapting item such as the Stimulus 7 /da/, the number of /ba/ responses to all members of the array tends to increase. Plotting only the number of /ba/ responses for each stimulus in the array (unlike the complementary plots in Figures 3 and 4), one sees that the crossover point, or 50 percent response level, has shifted toward the /da/ end of the continuum. In particular, Stimulus 4, which is normally identified as /ba/ on only about 40 percent of all trials in a preadaptation identification sequence, is now identified as /ba/ on better than 95 percent of all postadaptation identification trials.

At least two important aspects distinguish the adaptation effect with speech stimuli from the chromatic afterimage. First, it lasts a much longer time: the chromatic afterimage may last for only about 30 seconds or so, whereas the shift in the /ba/-/da/ identification function may last up to a few hours or even longer. This result is directly linked to the second difference between the phenomena. The chromatic afterimage does not transfer from one eye to another. That is, if one looks at a patch of blue with only the left eye open and then stares at a blank wall with only the right eye open, there will be no afterimage. This simple demonstration indicates that the locus of the chromatic effect is peripheral, or very near the retina and certainly before the neural pathways of the two eyes first converge in the lateral geniculate body. The adaptation effect with speech stimuli does transfer from one ear to the other and generally maintains its magnitude. This result indicates that the locus of the effect is central, and occurs after the pathways of the two ears converge (as low in the system as the superior olivary complex or as high as the cortex). These two factors, the duration of the effect and its locus, make it more similar to the visual work done in the 1960s by McCollough (1965) and by Blakemore and Campbell (1969), than to work with chromatic afterimages done originally in the nineteenth century. The adaptation work done in the field of speech perception, like the infant work discussed in Section D, was pioneered by Eimas (Eimas Cooper, and Corbit, 1973; Eimas and Corbit, 1973) and has been reviewed recently by Cooper (1975).

Several other aspects of speech adaptation are important and are closely related. First, the speech results have been interpreted in terms of feature adaptation. Features are thought to be processed by perceptual decision mechanisms. Adaptation shifts here contrast with the similar response shifts resulting from changes in adaptation level (Helson, 1964), since the latter may be accountable in terms of cognitive decision mechanisms. Second, these features are binary: that is, they are neurological correlates of the magical number two. Third, these features are often thought of as phonetic in nature, that is, as linguistic rather than as auditory. By extension, they have been thought unique to language. These three points deserve elaboration.

Feature analyzers as perceptual mechanisms. A major thrust of the first speech adaptation study (Eimas and Corbit, 1973) was that the apparent shifts in the identification functions were not attributable to response bias or other

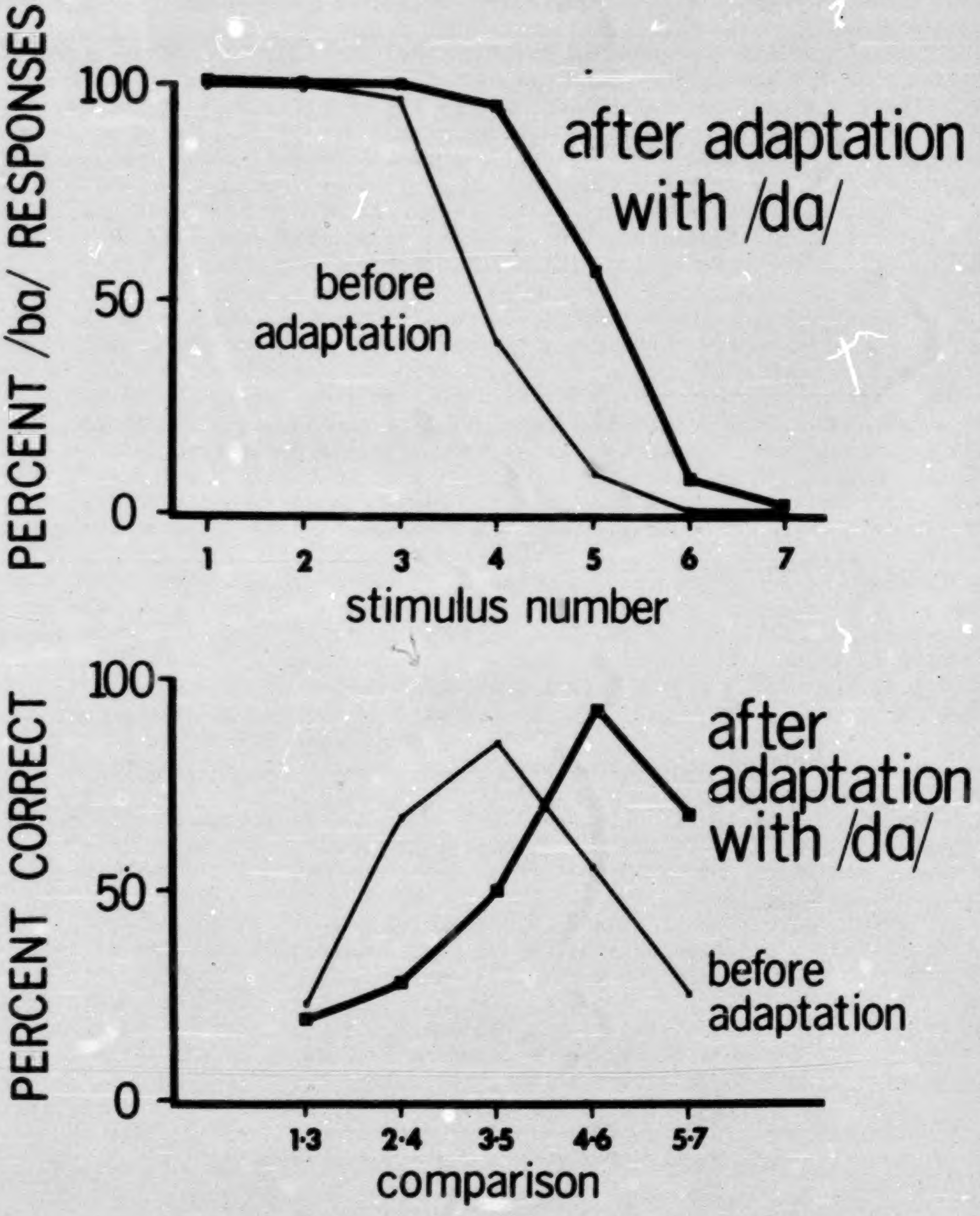

Figure 8: Schematic identification and discrimination data before and after adaptation with the Stimulus 7 /da/.

"conscious" shifts in decision criteria. As proof of this position one must find that not only do identification functions shift, but that corresponding discrimination functions also shift and by the same extent. Indeed, Eimas and Corbit (1973; and later Cooper, 1974) found that the discrimination functions do shift and by the anticipated amount.

Although their data did not deal with /ba/ and /da/, I may legitimately generalize as follows. In a preadaptation identification condition, Stimuli 1 through 3 are identified as /ba/ and Stimuli 5 through 7 as /da/. After adaptation with the Stimulus 7 /da/, Stimuli 1 through 4 are now identified as /ba/ and only Stimuli 6 and 7 as /da/. In other words, whereas Stimulus 4 is the most ambiguous item in the preadaptation condition, Stimulus 5 is most ambiguous in the postadaptation condition. If this change in identifiability is perceptual in nature, a preadaptation discrimination peak should be at the Stimulus-3/Stimulus-5 comparison, just as we have seen in Figure 4, whereas the postadaptation peak should be at the Stimulus-4/Stimulus-6 comparison. This is exactly the type of result found by Eimas and Corbit (1973), and is shown in the lower panel of Figure 8. These are schematic, not actual, data since these authors used different stimuli, but they accurately reflect their results. The categories of the magical number two change their locus with regard to the physical continuum, but they do not appear to change in any other manner.

It should be noted that postadaptation discrimination data are extremely difficult to gather. Only one or two discrimination trials are given after each long sequence of adapting stimuli. Thus, the task is very time consuming and only the most dedicated subjects will listen to the many hours of nonsense syllables, tediously and incessantly presented over and over again.

Feature analyzers as neural mechanisms. Underlying the shifts in identification and discrimination functions are some allegedly quite simple mechanisms. A scheme of how they might work is shown in Figure 9. Imagine two detectors in the perceptual system, one whose primary job it is to respond to the phoneme /b/ and the other to repsond to /d/. Each of these is maximally sensitive to a prototypic stimulus (perhaps Stimulus 1 for the /b/ detector and Stimulus 7 for the /d/ detector). In addition, each will respond to other neighboring stimuli as well, but at a somewhat reduced rate. The /b/ and /d/ detectors are relatively "close" to one another in that they can respond to the same stimulus, provided that it is roughly midway between the two stimulus prototypes along the physical continuum most relevant to the phonetic distinction, in this case the second-formant transition. Normally, the boundary between /b/ and /d/ is at the cross-over point of the sensitivity curves, as shown in the top panel. The Figure is drawn to be reminiscent of hypothetical signal detection functions and of a simplified one-dimensional rendition of Selfridge and Neisser's (1960) Pandemonium model of pattern recognition: at some neural level subsequent to the detectors themselves, a decision demon will decide which feature demon, that for /b/ or /d/, has yelled the loudest (which neuron has fired the most rapidly) and deserves to be recognized and identified over the Pandemonium of screams (neural firings) of all the other demons (feature analyzers). This ultimate decision determines the psychological identity of the stimulus.

During extensive adaptation to the same stimulus, repeated over and over, the particular feature analyzer may fatigue. The precise nature of the fatiguing process is not known, but one possibility is shown in the bottom panel of Figure 9. After adaptation with /da/, the /d/ analyzer may become less and less

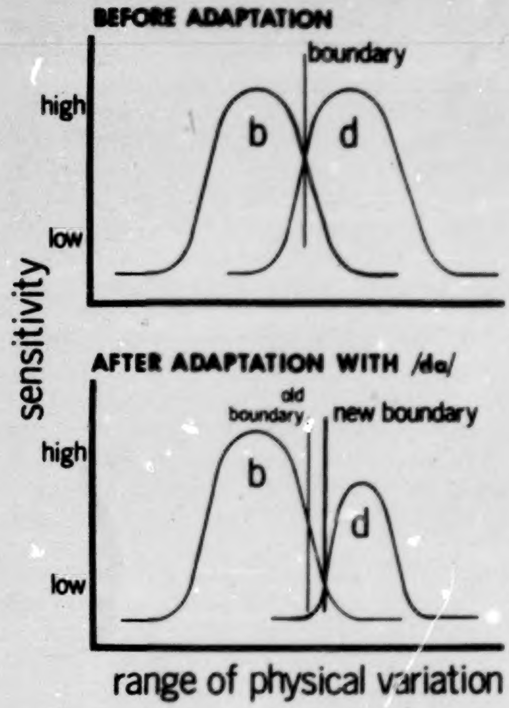

Figure 9: A simple binary-detector model to account for selective adaptation results.

sensitive to stimulation by the /d/ prototype and to all similar stimuli that would normally trigger it. The decrease in sensitivity may manifest itself in several ways. First the height of the sensitivity curve may decrease, then the shoulders of the function would slump inward. The effect would be that the new crossover point of the two feature analyzers would be moved slightly toward /d/, away from the old boundary. The new crossover would mark the locus of the post-adaptation boundary between /b/ and /d/. A refractory period of a considerable amount of time would be necessary to restore potency to the /d/ analyzer. Once restored, however, it would resume the sensitivity function shown in the top panel and consequently the old phoneme boundary would be restored as well.

This simple account seems to serve well in explaining shifts in identification and discrimination functions. Nevertheless, it is not without its problems. Consider one of the central aspects of this model: during adaptation, the feature analyzer becomes more and more insensitive to the prototype of the stimulus category. If this were true, we would expect to find some way to measure the decrement in sensitivity to the stimulus prototype. The data in identification functions provide no help. Each function necessarily asymptotes at 0 and at 100 percent by the time it reaches the prototype at Stimulus 1 or Stimulus 7. Hence we are restricted by floor and ceiling effects, and can make no inferences about any alleged attenuation in the sensitivity function. In a pilot study, Michael Posner and I supposed that a reliably more sensitive measure, that of reaction time, might serve to demonstrate the anticipated effect. Thus we adapted listeners to one of the endpoint stimuli in this /ba/-to-/da/ array, and measured their reaction time in responding to (identifying) each member of the array including the endpoint stimuli. Results seemed very clear. Indeed, there was a shift in the identification functions, just as we had anticipated, but there was no significant change in reaction time for the identification of stimulus prototype used in adaptation. An increase would have reflected, we thought, the decreased sensitivity to the category prototype; a decrease in reaction time might have reflected something else. No change is more difficult to interpret.

Perhaps, then, the sensitivity function shown in the bottom panel of Figure 9 is incorrect. Perhaps, rather than decreasing in "height," the function merely becomes more leptokurtic--that is, just as tall, but much slimmer. This would have the same effect in shifting the identification and discrimination functions. Regardless of the final form of the model, it probably does not differ greatly from the one presented here and should serve to account for all data. I shall return to this model later and discuss it in some detail from a different perspective. What must be considered now is the role of these feature analyzers in language and in music.

Feature analyzers as linguistic mechanisms and as auditory mechanisms. Linguists have talked of the binary features of language for a long time, and the influence of Jakobson, Fant, and Halle's book Preliminaries to Speech Analysis (1951) has been very influential in investigating the acoustic basis for these features. The notion of features was so well developed in the 1950s and 1960s that this term for various stimulus aspects of speech may have been borrowed by neurophysiologists when they discovered neural devices in the visual cortex that responded solely to edges moving in certain directions. The psychologist interested in speech and in effects of adaptation might look to linguists and find one use of the term features, look to neurophysiologists and find another use, and then yearn to close the gap between the two. Eimas and Corbit (1973) found this link and described it in their seminal paper. But were these detectors linguistic in nature, or not? There is a difference between a linguistic feature detector and a linguistically relevant feature detector. A linguistic feature detector would be one responding only to speech sounds and not to similar nonspeech sounds. A linguistically relevant feature detector, on the other hand, would respond to certain speech sounds but would also respond to relevant nonlinguistic sounds as well. As yet it is too soon to determine conclusively whether or not certain of these mechanisms are linguistic or merely linguistically relevant. Let me first present some evidence that supports the latter view, then evidence for the former.

A question arose, after the establishment and replication of the phenomenon, as to how linguistic these mechanisms are. That is to say, while there may be specific phonetic feature analyzers, are there general phonetic feature analyzers? The difference between specific and general is crucial. Specific detectors would be sensitive only to those aspects of a particular speech sound (/d/, for example) that occur in particular speech contexts (as in /da/). General detectors, however, would be sensitive to broader aspects of that speech sound /d/ as it occurs, for example, not only in /da/ but also in /di/, /du/, and /ad/ (as in deep, dupe, and odd). In other words, will the effect, as measured by shifts in identification functions pre- and postadaptation, transfer across different vowel environments? Will it transfer across different syllable positions? The answers are yes and no, respectively. Ades (1974) found that adaptation with /de/ (as in date) shifted the identification function of a /bae/-/dae/ continuum (as in bad, dad), but less readily than did adaptation with the endpoint /dae/ stimulus. He also found that adaptation with /dae/ had no effect on the identification of an /aeb/-/aed/ continuum. From such results the adaptation effect seems phonetic enough to transfer across small acoustic differences such as those seen when changing between similar vowel environments, but not phonetic enough to transfer across widely discrepant acoustic forms such as those found when shifting a phoneme from initial to final syllable position. Restated, the effect is general enough to transfer across small acoustic differences, but not general enough to transfer across large differences.

The fact that acoustic differences matter at all is an important issue: the postadaptation function for /bae/-/dae/ is shifted considerably by adaptation with /dae/, less by adaptation with /de/, and even less (that is, not at all) by adaptation with /aed/. Rather than account for this array of results by phonetic feature adaptation, one might account for the results in terms of acoustic or auditory feature adaptation. (I will use the term auditory so as to be more general than the term acoustic might allow. I will want to consider not only those acoustic features that are analyzed logically prior to the labeling of speech sounds but also entertain the possibility of a higher-level feature analysis of an auditory signal, which would appear to occur beyond the registration of the raw acoustic signal but would not necessarily involve language. Auditory seems the best term here, implying nonlinguistic as well as postsensory.) Before asserting that auditory feature analyzers are possible alternatives to phonetic feature analyzers, one must demonstrate that the notion of auditory analyzers is viable. Here, we have only to look to the pluck and bow sounds again.

We (Cutting, Rosner, and Foard, 1975) selected a 440-Hz (Concert A) sawtooth wave continuum to use for postadaptation identification. Items differed in rise time in 10-msec steps from 0 to 80 msec. Rather than just two, there were eight adaptation conditions: adaptation within the same continuum of sounds using the 0- and 80-msec rise time 440-Hz sawtooth items, which we take to be our "prototype" pluck and bow stimuli, plus six other conditions. There was adaptation across different frequencies using 0- and 80-msec sawtooth items at 294 Hz, adaptation across different waveforms using 0- and 80-msec sine wave items at 440 Hz, and adaptation across both frequency and waveform using 0- and 80-msec sine wave items at 294 Hz. The same very diligent listeners served in all conditions but on eight separate days, one per condition. A preadaptation identification test was given before adaptation tests, and comparisons were always made between pre- and postadaptation identification functions for a particular day.

Results were very clear in support of auditory analyzers of pluck and bow. Adaptation within the 440-Hz sawtooth wave continuum was considerable. The normal boundary of about 40-msec rise time shifted to about 37-msec rise time in the pluck-adaptation condition, and to about 50-msec rise time in the bow-adaptation condition. Both shifts were highly significant and the difference in their size is common in adaptation findings. Such differences are often difficult to interpret, but here we think that they relate to inherent limitations in a continuum such as rise time. For example, a stimulus can be no more abrupt than 0-msec rise time, but can be infinitely less abrupt than 80 msec.

Significant postadaptation boundary shifts occurred in nearly all other conditions as well, but their magnitudes tended to be smaller. For example, pluck-adaptation shifts across one dimension (frequency or rise time) averaged less than 3 msec, and complementary bow-adaptation shifts averaged only 5 msec.

Adaptation boundary shifts across both dimensions (frequency and rise time) were smaller still: only 2 msec for both pluck and bow. These results are regular enough to allow a simple interpretation: the more stimulus dimensions held in common between adapting and test stimuli the larger is the effect. Thus these results for pluck and bow items are very similar to the array of results for speech items.

Let us suppose that the somewhat smaller adaptation effect of /de/ than /dae/ on a /bae/-/dae/ test continuum is due to the fact that fewer dimensions are shared between /de/ and the test items. These dimensions could be linguistic (the vowels differ between adapting and test stimuli) or they could be auditory (the spectra differ as well). That there is no adaptation effect of /aed/ on /bae/-/dae/ is difficult to account for in linguistic terms without relying on some allophonic or syllabic level of processing; but the result is relatively easy to account for in auditory terms because the stimuli are very different. Ignoring the common vowel nuclei, the comparable transitions in all formants go the wrong way, that is, in opposite directions. For /dae/ the first formant ascends in frequency into the following vowel, whereas for /aed/ it descends, following it. The reverse is true for the second formants. Perhaps, then, all shifts due to adaptation are actually auditory in nature rather than linguistic: perhaps they are due to adaptation of linguistically relevant features, not linguistic features.

How, then, does adaptation actually work? Perhaps it is not the feature analyzer itself that is fatigued. Instead, it may be the neural pathway leading to the analyzer that suffers fatigue. The more the adapting stimulus differs from the test-item array, the fewer may be the number of intervening processing stages (from the registration of the acoustic signal to the binary feature analyzers) that are shared between adapting and test stimuli. Thus seen, fatigue during adaptation might be an inhibitory process that builds up throughout the many neurons and synapses of a particular pathway for a particular signal. If this notion is correct, then when the test array differs from the adapting stimulus, each item in that array will travel a somewhat different neural path from that of the adapting stimulus. The sections of the path for the test stimuli that are not held in common with the adapting item will not be fatigued, while those portions held in common will be fatigued. Roughly speaking, if only half of the pathway neurons are held in common, perhaps the adaptation effect might be only half as great. It should be obvious that the use of the term pathway here is at least partly metaphorical, but I do not mean it to be exclusively so (see Posner, 1975).

If the pathway-fatigue account of selective adaptation is viable, the sensitivity functions shown in Figure 9 and the decrease shown in one function in the bottom panel of that figure, may not mark the sensitivity of the actual analyzers themselves. Instead, they may mark, for a given stimulus, the rate of firing of the sequence of particular neurons in the pathway that leads to the analyzer: adaptation leads to pathway fatigue and less neural activity. The end result is exactly the same. Rather than plot sensitivity on the ordinate of that function, one might substitute the term "neural activity." Boundary shifts would occur in the same manner.

What about the notion of linguistic feature analyzers as opposed to linguistically relevant analyzers? The answer can come only after exploring further the notion of pathways. It seems to make no sense to think of a linguistic pathway; that is, a neural route that is traveled exclusively by speech stimuli. Such a pathway would necessarily have to have an early gating device that has already decided that a particular stimulus is speech. If the speech/nonspeech decision were already made at this early level, then subsequent analyses would appear to be unnecessary: to determine whether or not an item is speech, the system would surely have analyzed the signal for speechlike features. Instead, then, the pathways leading to the binary analyzers are auditory; that is, general enough to handle both linguistic and nonlinguistic events. Since seemingly all the variance in boundary shifts across different types of adaptation situations might be accounted for by pathways, the existence of linguistic analyzers would be unimpugned. It is clear that there are auditory analyzers of similar nature, at least for pluck and for bow, but they may exist side-by-side, as it were, with linguistic analyzers. The only logical argument I can offer against this possibility is an appeal to parsimony: Why have two kinds of binary feature analyzers, linguistic and nonlinguistic, when one set of nonlinguistic analyzers might do? It may be that the assignment of particular speech labels, such as /b/ and /d/, occur subsequent to their categorization. Unfortunately, such speculation takes us uncomfortably far from the available data.

As an internal summary then, postadaptation boundary shifts in identification functions of certain speech and musiclike stimuli seem to be explicable in terms of neural fatigue. Exactly which pathways fatigue remains an important question. I have suggested that the fatigue takes place in the neural pathways prior to arrival at the feature analyzers. Such an account would easily allow for differential magnitudes of boundary shifting according to the number of differences between adapting and test stimuli. The analyzers themselves lie considerally beyond the point at which the two ears converge, and they may not suffer from adaptation "fatigue." They appear to be binary, at least with regard to any one stimulus, and they appear to function according to signal-detection criteria and a simplified Pandemonium model. It seems clear that they can be either linguistic or nonlinguistic in nature. Most may be solely linguistic and the few others may be linguistically relevant. Remember that rise time cues not only the difference between pluck and bow, but also the difference between /ʃa/ and /tʃa/. At present, we have found only one set of analyzers that overlaps the domains of speech and music. It would be of considerable interest if others can be found that perform this apparent dual function. It would also be of interest to find possible binary music analyzers that are not relevant to speech.

F. Summarizing Remarks

What of the magical number two? Part of the answer is the same as for the magical number seven. Miller (1956) noted seven wonders of the world, seven points on a psychological rating scale, seven seas, seven categories of absolute judgment, seven deadly sins, seven objects in the span of attention, seven days of the week, and seven digits in the span of immediate memory. He suspected that all these sevens were merely "pernicious, Pythagorean coincidence." Such coincidence is found even more easily in twos: the two faces of Janus; the two types of learning, operant and respondent; the two cosmic forces, Yin and Yang; the two minds, conscious and unconscious; the two sexes, male and female, the two searches through memory, self-terminating and exhaustive; the two diurnal segments, day and night; the two locales for research, laboratory and field;

and many other twos, such as up and down, self and others, and indeed /ba/ and /da/, and pluck and bow. Jakobson, Fant, and Halle suggested that the dichotomous scale is the pivotal principle of linguistic structure; a quick glance at twoness elsewhere suggests that it may be the pivotal principle by which we parse the world (see Ogden, 1932).

Beyond any "perniciousness" of the magical number two, humans appear to be predisposed to perceive certain auditory events in a dichotomous manner. I first discussed such discrete perception in terms of identification functions for an acoustic continuum of speech sounds from /ba/ to /da/ generated by a computer-driven speech synthesizer. This first expression of the magical number two proves not to be crucial. Lines slanted at various angles yield equally quantal identification functions. The crux of the magical number two is revealed in the discrimination functions, where for the stop consonants one can discriminate only as well as one can identify, but for the slanted lines one can discriminate almost infinitely better. A continuum of vowel sounds from /i/ to /ɪ/ yields intermediate results, and these results appear to be interpretable in terms of a small perceptual deviation from the number two. The peculiar nonlinearity found in the discrimination of stop-consonant sounds is not unique to speech items. In fact, evidence for the dichotomous perception of musiclike sounds is just as striking as that for speech.

Our perceptual predisposition toward the magical number two appears to stem from biological endowment. Infants as young as one and two months parse certain speech and music sounds in a manner functionally identical to that of adults. Such results indicate that these categories and boundaries are natural according to the most stringent criterion; indeed, they appear to be innate.

The neural mechanisms underlying our perception by twos can be thought of as yoked-pairs of feature analyzers lying well beyond the cochlea. Some may be unique to speech analysis, others, like those for pluck and bow musical sounds, may be used in both speech and nonspeech analysis. Continued presentation of a particular stimulus prototype appears to fatigue selectively the feature-analysis system, and temporary shifts in the locus of category boundaries are obtained. These mechanisms appear to allow for high-speed speech perception, the rapid categorization of a particular speech sound, and the discarding of its nonprototypic vagaries; and they also allow for great savings, since the speech sounds are coded into tight bundles of phonetic features suitable for memory and storage.

In music, however, the role of feature analyzers is less clear. Pluck and bow categories obviously discriminate among modes of playing certain musical instruments, and may relate to the reason these various stringed instruments were invented, but beyond this minor role and in the absence of other binary musical features (which may yet be discovered), we simply do not know why they exist. It may be that they were auditory precursors to phonetic feature detectors, as if nature were experimenting with the feasibility of such devices. It may be that they are related to orienting and startle mechanisms: sounds with rapid onsets often forebode danger, whereas sounds with more gradual onsets are more likely to be associated with "safer" events. Beyond these speculations it is too early to say what their purpose may be. Indeed, as Miller suggests, the mere existence of pluck and bow categories may be a "pernicious, Pythagorean coincidence," but it is preferable to think that they will eventually be tied to a theoretical fabric relating the structures of speech and music.

REFERENCES

Ades, A. E. (1974) How phonetic is selective adaptation? Experiments on syllable position and vowel environment. Percept. Psychophys. 16, 61-66.

Barclay, J. R. (1972) Noncategorical perception of a voiced stop: A replication. Percept. Psychophys. 11, 269-273.

Blakemore, C. and F. W. Campbell. (1969) On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. J. Physiol. 203, 237-260.

Cooper, W. E. (1974) Adaptation of phonetic feature analyzers for place of articulation. J. Acoust. Soc. Amer. 56, 617-627.

Cooper, W. E. (1975) Selective adaptation to speech. In Cognitive Theory, Vol. 1, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman, and D. B. Pisoni. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Cutting, J. E. and P. D. Eimas. (1975) Phonetic feature analyzers and the processing of speech by infants. In The Role of Speech in Language, ed. by J. F. Kavanagh and J. E. Cutting. (Cambridge, Mass.: MIT Press).

Cutting, J. E. and B. S. Rosner. (1974) Categories and boundaries in speech and music. Percept. Psychophys. 16, 564-570.

Cutting, J. E., B. S. Rosner, and C. F. Foard. (1975) Rise time in nonlinguistic sounds and models of speech perception. Haskins Laboratories Status Report on Speech Research SR-41, 71-94.

Eimas, P. D. (1974) Auditory and linguistic processing of cues for place of articulation by infants. Percept. Psychophys. 16, 513-521.

Eimas, P. D., W. E. Cooper, and J. D. Corbit. (1973) Some properties of linguistic feature detectors. Percept. Psychophys. 13, 247-252.

Eimas, P. D. and J. D. Corbit. (1973) Selective adaptation of linguistic feature detectors. Cog. Psychol. 4, 99-109.

Eimas, P. D., E. R. Siqueland, P. Jusczyk, and J. Vigorito. (1971) Speech perception in infants. Science 171, 303-306.

Helson, H. (1964) Adaptation-Level Theory: An Experimental and Systematic Approach to Behavior. (New York: Harper & Row).

Hockett, C. F. (1960) Logical considerations in the study of animal communication. In Animal Sounds and Communication, ed. by W. E. Lanyon and W. N. Tavolga. (Washington, D.C.: American Institute of Biological Sciences).

Jakobson, R., C. G. M. Fant, and M. Halle. (1951) Preliminaries to Speech Analysis. (Cambridge, Mass.: MIT Press, 1963).

Jusczyk, P. W., B. S. Rosner, J. E. Cutting, C. F. Foard, and L. Smith. (1975) Categorical perception of nonspeech sounds in the two-month-old infant. Paper presented to the Society for Research in Child Development, Denver, Col.

Lane, H. (1965) Motor theory of speech perception: A critical review. Psychol. Rev. 72, 275-309.

Lane, H. (1967) A behavioral basis for the polarity principle in linguistics. Language 43, 494-511.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington, D.C.: V. H. Winston & Sons).

Locke, S. and L. Kellar. (1973) Categorical perception in a nonlinguistic mode. Cortex 9, 355-369.

McCollough, C. (1965) Color adaptation of edge-detectors in the human visual system. Science 149, 1115-1116.

Miller, G. A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol. Rev. 63, 81-97.

Morse, P. A. (1972) The discrimination of speech and nonspeech stimuli in early infancy. J. Exp. Child Psychol. 14, 477-492.

Morse, P. A. and C. T. Snowdon. (1975) An investigation of categorical speech discrimination by rhesus monkeys. Percept. Psychophys. 17, 9-16.

Ogden, C. K. (1932) Opposition. (Bloomington: Indiana University Press, 1967).

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Unpublished Ph.D. dissertation, University of Michigan.

Pisoni, D. B. (1973) Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253-260.

Pisoni, D. B. (1975) Auditory short-term memory and vowel perception. Mem. Cog. 3, 7-18.

Pisoni, D. B. and J. H. Lazarus. (1974) Categorical and noncategorical modes of speech perception along the voicing continuum. J. Acoust. Soc. Amer. 55, 328-333.

Pisoni, D. B. and J. Tash. (1974) Reaction times to comparisons within and across phonetic categories. Percept. Psychophys. 15, 285-290.

Posner, M. I. (1975) The temporal course of pattern recognition in the human brain. In Signal Analysis and Pattern Recognition in Biomedical Engineering, ed. by G. F. Inbar. (Tel Aviv: Israel Universities Press).

Potter, R. K., G. A. Kopp, and H. C. Green. (1947) Visible Speech. (New York: Van Nostrand).

Rosch, E. H. (1973) Natural categories. Cog. Psychol. 4, 328-350.

Selfridge, O. and U. Neisser. (1960) Pattern recognition by machine. Sci. Amer. 203 (Aug.), 60-68.

Studdert-Kennedy, M., A. M. Liberman, K. S. Harris, and F. S. Cooper. (1970) Motor theory of speech perception: A reply to Lane's critical review. Psychol. Rev. 77, 234-249.

Processing Two Dimensions of Nonspeech Stimuli: The Auditory-Phonetic Distinction Reconsidered*

Mark J. Blechner,+ Ruth S. Day,+ and James E. Cutting++

ABSTRACT

Nonspeech stimuli were varied along two dimensions: intensity and rise time. In a series of speeded classification tasks, subjects were asked to identify the stimuli in terms of one of these dimensions. Identification time for the dimension of rise time increased when there was irrelevant variation in intensity; however, identification of intensity was unaffected by irrelevant variation in rise time. When the two dimensions varied redundantly, identification time decreased. This pattern of results is virtually identical to that obtained previously for stimuli that vary along a linguistic and a nonlinguistic dimension. The present data, taken together with those of other studies using the same stimuli, suggest that the mechanisms underlying the auditory-phonetic distinction should be reconsidered. The results are also discussed in terms of general models of multidimensional information processing.

Several contemporary accounts of speech perception have emphasized the organization of processing into a hierarchy of levels, including auditory, phonetic, phonological, lexical, syntactic, and semantic (Fry, 1956; Stevens and House, 1972; Studdert-Kennedy, in press). The distinction between phonetic and higher levels has been commonly accepted by linguists and psychologists for some time. Recently, however, much attention has been directed toward the auditory-phonetic distinction (e.g., Fant, 1967; Stevens and Halle, 1967; Studdert-Kennedy, Shankweiler, and Pisoni, 1972). Fry (1956), in an early discussion of the levels-of-processing view of speech perception, emphasized the role of the "physical-psychological transformation" that occurs in the recognition of phonemes from the acoustical signal. The important characteristic of this transformation is that there is no one-to-one relationship between "the number and arrangement of physical clues and the sound which is recognized" (p. 170).

---

*A revised version of this paper will appear in Journal of Experimental Psychology: Human Perception and Performance.

+Also Yale University, New Haven, Conn.

++Also Wesleyan University, Middletown, Conn.

Acknowledgment: This research was supported in part by NIMH Training Grant PHS5T01MH05276-27 to Yale University. The authors thank Michael Studdert-Kennedy for a discussion of the issues raised in this paper, and Robert L. Plotz for assistance in running the experiment.

Fry did not state that the physical-psychological transformations characteristic of speech are exclusive to speech. However, this possibility was emphasized by later work that viewed speech perception as mediated by articulatory mechanisms (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Stevens and House, 1972). Such a view made it desirable and perhaps necessary to partition all sounds into two general classes: those that are speech and those that are not.[1]

Definitions concerning which criteria must be met in order for sounds to be classified as speech have varied in the literature. The present paper assumes a two-part definition of speech: sounds that (1) can be articulated by the human vocal apparatus; and (2) can be recoded into higher-order linguistic units. According to this definition, phonetic processing may make reference to articulatory processes either directly (Liberman et al., 1967) or implicitly (Stevens and House, 1972), and follows some system of linguistic organization such as a distinctive feature system (Jakobson, Fant, and Halle, 1963; Chomsky and Halle, 1968). Furthermore, while all sounds undergo auditory processing, only speech sounds undergo phonetic processing.

To determine the empirical validity of the auditory-phonetic distinction, many experiments have been conducted. Results from several paradigms suggest that even though speech sounds differ along a wide variety of acoustic dimensions, they are perceived in ways that are qualitatively distinct from the way nonspeech sounds are perceived. For example, the main difference between the phonemes /ba/ and /da/ is the direction and extent of the second-formant transition (Liberman, Delattre, Cooper, and Gerstman, 1954), while the distinction between /ba/ and /pa/ lies in the latency between the initial plosive burst and the onset of voicing (Lisker and Abramson, 1964). Yet, for both distinctions, there is no one-to-one relationship between changes in the acoustic patterns and probabilities of identification. Instead, item identifications remain at or near 100 percent as one phoneme or the other, with an abrupt crossover at the "phoneme boundary." More importantly, whereas most stimulus dimensions in the environment, such as pitch, intensity, and brightness, can be discriminated from another much more accurately than they can be identified (Miller, 1956), this is not the case for several linguistic dimensions. Instead, two acoustically different stimuli that lie within the same phoneme category are discriminated at near-chance level; two stimuli that lie in separate categories but differ by the same acoustic increment are discriminated with very few errors. This nonlinear mode of perception is called "categorical perception" (Liberman, Harris, Hoffman, and Griffith, 1957; for a review, see Studdert-Kennedy, in press).

Other experimental operations have shown processing differences for speech and nonspeech stimuli. Recent work using a selective adaptation paradigm has shown that repeated presentation of a consonant-vowel (CV) syllable produces systematic shifts in the phoneme boundary (Eimas, Cooper, and Corbit, 1973; Eimas and Corbit, 1973). Some experiments suggest that the basis of this adaptation is phonetic rather than auditory (Cooper, 1975), although this evidence is not conclusive (Studdert-Kennedy, in press). The auditory-phonetic distinction seems to be further supported by dichotic identification tasks that often reveal right-ear advantages for speech stimuli (e.g., Kimura, 1961; Shankweiler and Studdert-Kennedy, 1967) and left-ear advantages for nonspeech stimuli (e.g., Kimura, 1964; Chaney and Webster, 1966; Curry, 1967).

---

[1] Some authors use the terms "linguistic and nonlinguistic" or "verbal and nonverbal" to indicate the same distinction.

In addition, several experiments have investigated the relationship of auditory and phonetic processes in selective attention tasks using stimuli that vary along two dimensions. When both dimensions are linguistic (e.g., initial stop consonant and vowel in CV syllables), selective attention for either dimension is impaired by irrelevant variation in the other dimension (Wood and Day, 1975). This pattern of symmetric interference also occurs when both dimensions are nonlinguistic, such as pitch and intensity (Wood, 1975). However, when one dimension is linguistic and the other is not, a pattern of asymmetric interference appears; reaction time (RT) for identification of stop consonants is impaired by irrelevant variation in pitch, but RT for pitch identification is not increased significantly by irrelevant variation in stop consonant (Day and Wood, 1972; Wood, 1974, 1975). These results have been interpreted to support the auditory-phonetic distinction.

The dichotic listening, categorical perception, selective adaptation, and speeded classification experiments appear to comprise a set of converging operations (Garner, Hake, and Eriksen, 1956) on the psychological reality of the auditory-phonetic distinction. Recently, however, this view has been brought into question by experiments in which the stimuli are sawtooth-wave tones differing in rise time (see Cutting, in press). While rise time can cue a speech distinction, such as the difference between the /ʃa/ and /tʃa/, sawtooth waves are not perceived as speech. Instead, they sound comparable to a plucked or bowed violin string. Although these "plucks" and "bows" are not processed phonetically, they are processed similarly to speech in several ways. They are perceived categorically (Cutting and Rosner, 1974), and their identification functions shift in the same manner as for speech following selective adaptation (Cutting, Rosner, and Foard, 1975). Ear advantage data for plucks and bows are not yet decisive.

The present experiment seeks to investigate further the processes by which the plucks and bows are perceived and their relationship to the auditory-phonetic distinction. The two-choice speeded classification procedure developed by Garner and Felfoldy (1970) and modified for use in auditory experiments by Day and Wood (1972; Wood, 1974, 1975; Wood and Day, 1975) was used to determine how the dimensions of rise time and intensity interact. If symmetric interference necessarily results when stimulus dimensions are of the same general class--i.e., both linguistic or nonlinguistic--then selective attention to either rise time or intensity should suffer from irrelevant variation in the other dimension. However, if a pattern of asymmetric interference occurs, it would be clear that such a pattern need not be based on separate auditory and phonetic levels of processing. This pattern of results, along with those of other studies using pluck and bow stimuli, would lead one to question the mechanisms underlying the auditory-phonetic distinction as currently conceived.

The present experiment is also concerned with current conceptions about the processing of multidimensional stimuli in general. The pattern of asymmetric interference suggests a serial model; only the dimension processed first interferes with the other. However, this view has been challenged by the finding that when the two dimensions of the same stimulus vary redundantly, subjects can identify them more quickly than when only one dimension varies (Wood, 1974). This redundancy gain (Garner and Felfoldy, 1970) seems to argue against a serial model. If pitch processing were completed before processing of the stop consonant began, a subject should not be able to use redundant information about the stop consonant to speed up identification of pitch. Instead, a parallel model, which posits that processing of the two dimensions overlap or are simultaneous, would better account for the redundancy gain. Information-processing models that propose strict serial or parallel processing do not seem to offer an adequate explanation of Wood's (1974) finding of asymmetric interference with redundancy gain. It seems likely, instead, that subjects have some degree of freedom about the kinds of processing that they use in different task conditions. The present experiment, by also including a task that varies the two dimensions redundantly, seeks to distinguish the conditions in which processing strategies are optional from those in which they are mandatory.

METHOD

Stimuli

Stimuli varied along two dimensions--intensity and rise time. They were derived from the sawtooth waves used by Cutting and Rosner (1974), generated on the Moog synthesizer at the Presser Electronic Studio of the University of Pennsylvania. The original stimuli differed in rise time and resembled the sound of a plucked or bowed violin string. The pluck and bow stimuli reached maximum intensity in 10 and 80 msec, respectively. Two more stimuli of lower amplitude were created by attenuating the original stimuli 7 dB, using the pulse code modulation (PCM) system at the Haskins Laboratories (Cooper and Mattingly, 1969). Thus, the final four stimuli were loud-pluck, soft-pluck, loud-bow, and soft-bow. The absolute level of the loud and soft stimuli were 75 and 68 dB re 20 $\mu N/m^2$, respectively. All stimuli were truncated to 800 msec in duration (the original stimuli were approximately 1050 msec), and then digitized and stored on disc file using the PCM system. Items were reconverted to analog form at the time of tape recording.

Tapes

All tapes were prepared on the PCM system. Test stimuli were recorded on one channel of the audio tape. On the other channel, brief pulses were synchronized with the onset of each test stimulus; these pulses triggered the RT counter during the experimental session.

A display tape was prepared to introduce the listeners to the stimuli. The four stimuli were played in the same order several times, beginning with three tokens of each item, then two of each, and finally one of each. Practice tapes were also prepared, consisting of a randomized order of 20 items, five of each stimulus. There were two practice tapes, each with a different random order.

The eight test tapes each contained 64 stimuli with a 2-sec interstimulus interval. Each tape was composed of different subsets of the four stimuli, depending on the condition of the experiment. In the control condition, the stimuli varied along only one dimension, while the other dimension was held constant. Thus, for example, one intensity control tape consisted of loud and soft bows only, while the other consisted of loud and soft plucks. For half the subjects, the nontarget dimension (in this case, rise time) was held constant at one value (pluck), whereas for the other half, it was held constant at the other (bow). Rise-time control tapes were constructed in an analogous fashion. In the orthogonal condition, both dimensions varied independently. Hence, the two tapes for this condition contained all four kinds of stimuli, in different random orders.

In the correlated condition, the dimensions varied in a completely redundant manner. In this experiment all of the pluck stimuli in this condition were loud and all of the bow stimuli were soft. See Table 1 for a complete outline of the stimuli in each condition.

TABLE 1: Stimulus sets for each target dimension and condition.

	Condition		
Target Dimension	Control	Correlated	Orthogonal
Rise Time	Loud-pluck, loud-bow or Soft-pluck, soft-bow	Loud-pluck Soft-bow	Loud-pluck Soft-pluck Loud-bow Soft-bow
Intensity	Loud-pluck, soft-pluck or Loud-bow, soft-bow	Loud-pluck Soft-bow	Loud-pluck Soft-pluck Loud-bow Soft-bow

Subjects and Apparatus

The six subjects (five males and one female, from 19 to 27 years of age) participated in all six tasks. All reported no history of hearing trouble.

The tapes were played on an Ampex AG-500 tape recorder and the stimuli were presented through calibrated Telephonics headphones (Model TDH39-300Z). Subjects sat in a sound-insulated room and responded with their dominant hand on the two telegraph keys mounted on a wooden board. Throughout the experiment, the left key was used for pluck and loud responses, while the right key was used for bow and soft responses. The pulse on one channel of the tape triggered a Hewlett-Packard 522B electronic counter. When a response on either telegraph key stopped the counter, the reaction time was registered onto paper tape by a Hewlett-Packard 560A digital recorder for subsequent analysis. The listener's response choice was recorded manually by the experimenter.

Procedure

At the start of the session, subjects were informed of the general nature of the experiment and of the dimensions they would be asked to identify. They were told that the difference in rise time could be compared to the difference in sound between a plucked and a bowed violin string.

For preliminary training, subjects heard the display sequence twice. Next, they listened to the two sets of 20 practice items and responded verbally, first attending to rise time and then to intensity. This ensured that the subject could perceive the differences along both dimensions. They then repeated the same practice trials, responding with a key-press rather than verbally, in order to become familiar with the mode of response. Subjects were instructed to respond as quickly as possible without making errors.

The order of presentation for the six test tapes was determined by a balanced Latin Square design. Before the test trials of each condition, appropriate instructions were read. The subject was then given eight practice trials to help stabilize RT performance and to familiarize him with the identification task and stimulus set for that particular condition. In the control conditions the subject was told which dimension to attend to and the value at which the other dimension would be held. In the orthogonal conditions, he was instructed to attend to one dimension and to ignore variation in the other dimension. In the correlated conditions, he was instructed to attend to one dimension, but was encouraged to use the additional information from the other dimension.

RESULTS

Both dimensions were easy to identify. In the practice trials, no subject made more than 2 percent errors. During test trials, the highest mean error rate for any condition was 1.8 percent. A three-way analysis of variance of the error data (subjects × conditions × dimensions) revealed no significant main effects nor interactions. Therefore, the error data will not be considered in any detail in this discussion.

For the reaction time data, median RT was calculated for each individual block of trials for each subject, and means of medians for each condition across subjects were computed.[2] In addition, the untransformed RT data were subjected to a complete four-way factorial analysis of variance (subjects × conditions × dimensions × within cell).

Median RT data are presented in Table 2. For the dimension of rise time, there was an increase of 53.5 msec from the control to the orthogonal condition, while for the dimension of intensity, there was an increase of only .2 msec. In the analysis of variance, the effect of dimensions (intensity versus rise time) was not significant, while the effect of conditions (control, orthogonal, and correlated) was significant [F (2, 126) = 33.23, $p<.001$]. In addition, the dimensions × conditions interaction was significant [F (2, 126) = 6.43, $p<.01$]. In order to differentiate interference effects from redundancy gains, a contrast of the interactions between the two dimensions and only the control and orthogonal conditions was performed, omitting correlated conditions. This contrast was significant [F (1, 63) = 16.58, $p<.001$]. Thus there was an asymmetric pattern of interference; intensity variation interfered with the processing of rise time, while rise time had virtually no effect on the processing of intensity. This finding is especially interesting given that intensity was somewhat more discriminable than rise time, although the difference of 14.2 msec between the control conditions was not statistically reliable.

---

[2] Wood and Day (1975), in all of their reaction-time experiments, transformed their data, so that any RT longer than 1 sec was set equal to 1 sec. This was done to correct for possible malfunctioning of the equipment, such as failure of the response key to make electrical contact, or temporary inattention of the subject. While we agree that unusually long reaction times due to equipment trouble or lapsing of the subject's attention should be transformed, the continual resetting of long RT values to 1 sec seems arbitrary and could distort the data. If arithmetic means are to be used, very long data points can be more equitably adjusted by using reciprocal RT values. Alternatively, medians or trimeans can be used, with similar effect.

TABLE 2: Median reaction time in milliseconds for each dimension and condition.

		Condition	
Dimension	Control	Correlated	Orthogonal
Rise time	426.5	393.7	480
Intensity	442.3	406.4	442.5

The effect of redundancy gains was assessed by two methods. A contrast of the conditions effect showed the correlated conditions to be significantly different from the control conditions [F (1, 63) = 35.1, p<.001]. A subsequent comparison of the individual means using the Newman-Keuls procedure showed the correlated conditions in both dimensions to differ significantly from the respective control condition. The different correlated conditions, like the control conditions, did not differ significantly from each other.

In order to determine whether the redundancy gain could rightfully be considered as evidence of parallel processing of the two dimensions in this experiment, three alternative explanations of the redundancy gain were ruled out, as in Wood (1974). First, the possibility of a different speed-accuracy trade-off in the two correlated conditions could be eliminated by the lack of significant differences in the error data, as noted above.

Second, the possibility that the redundancy gain could be due to selective serial processing (SSP; see Felfoldy and Garner, 1971) was considered. If a subject uses the SSP strategy in the correlated condition, he merely attends to the more discriminable of the two dimensions for him, regardless of the instructions. Thus, his RT data would show that neither correlated condition is faster than the faster control condition. To test for the occurrence of the SSP strategy, the RT data for each correlated condition was tested against each subject's faster control condition with an analysis of variance and a subsequent comparison of means using the Newman-Keuls method. The correlated conditions were still found to be faster than the faster control condition [F (2, 63) = 13.5, p<.001]. Therefore, the redundancy gain cannot be attributed to the SSP strategy.

Finally, because the stimulus sets in the two correlated conditions were the same, whereas in the control conditions they were different, it is possible that the redundancy gain could be based on differential transfer between control and correlated conditions (Biederman and Checkosky, 1970). To test this explanation, an analysis of variance of the control and correlated conditions (subjects × conditions × order of presentation × within cell) was performed. The control condition presented first was 16 msec slower than the second, suggesting a possible practice effect, although this difference was not significant. The correlated conditions presented first and second differed by only 2 msec--which was not significant. Thus the transfer between the correlated conditions was less than or equal to the transfer between the control conditions, so that differential transfer does not account for the redundancy gain.

DISCUSSION

The pattern of results in this experiment is remarkably similar to those of Wood (1974). The relationship of intensity to rise time matches that of pitch to initial stop consonant both in the asymmetric pattern of interference and in the significant redundancy gain. In Garner's (1974) terminology, the dimensions of rise time and intensity are therefore asymmetrically integral. In the following discussion we will consider first the implications of our results in terms of general information-processing models and then reconsider the auditory-phonetic distinction.

The present results pose problems for information-processing models that try to account for perception only in terms of the serial or parallel handling of information. The data suggest that both stimulus and task characteristics may affect the mode of processing, so that neither a strict serial nor a strict parallel model can account for the whole picture. This view agrees with the recent suggestions of several authors (Nickerson, 1971; Townsend, 1971; Garner, 1974). An alternative to the strictly serial and parallel models is that the two processes overlap temporally and that one is contingent on the other, as suggested by Turvey (1973). Perhaps the processing of rise time and intensity (as well as of place of articulation and pitch) begin simultaneously. Both kinds of information can be combined to produce a redundancy gain in the correlated condition. However, it may be that only the orthogonal condition, which requires information gating (Posner, 1964), can reveal the contingency of one kind of information on the other. Current theories, however, do not account for why this contingency relationship might affect one task and not the other. Further research is needed on this point.

One useful approach might be to vary the discriminability of the two dimensions and to note the changes that occur in both the orthogonal and correlated conditions. For example, Blechner and Cutting[3] performed a speeded classification experiment using rise time and pitch, where the latter dimension was considerably more discriminable than the former. The result was that the subjects always processed the more discriminable dimension first. Thus, in the orthogonal condition, RT performance was equal to the faster control condition (the SSP pattern). However, it would be more interesting to determine whether, by manipulating discriminability in the reverse direction, rise time can be made to interfere with intensity. How much more discriminable than intensity would rise time have to be for a pattern of mutual interference to appear? If RT performance in the correlated condition were only as fast as the faster control condition, one might conclude that redundancy gains are impervious to the effects of contingency processing relationships between dimensions. Such a finding would be congruent with the results of the experiment reported here.

The present results bear not only on the way that different levels of processing interact, but also on the very question of which levels of processing are important in human auditory perception. In light of the present data, which show asymmetric interference between two "auditory" dimensions, it seems unwise to lump together all acoustic properties that do not provide linguistic cues. At the very least, a heterogeneous conception of auditory analyzing systems, with processing levels of increasing complexity, seems preferable.

---

[3] M. J. Blechner and J. E. Cutting. (in preparation) Selective serial processing of two-dimensional auditory stimuli.

The present data also suggest several basic positions concerning the issue of speech and nonspeech processing systems. One view suggests that two analogous processing systems exist--one for speech and one for nonspeech. Both are subsequent to a preliminary analysis of all acoustic stimuli, and both are organized to produce identical results in the speeded classification task. This explanation, however, may be challenged on grounds of parsimony: Why conceive of two systems when both behave in the same way in several circumstances? Cautious theorists, however, may object that analogy is not identity, and that the theory of separate speech and nonspeech systems should remain as long as perfect congruence between the two systems has not been demonstrated.

An alternative account of the present results is that perceptual processes are not divided according to the status of the stimulus dimensions as speech or nonspeech, but rather with respect to the kinds of acoustic analysis that the signal requires. Thus, rise time seems to be perceived in the same way, regardless of whether it characterizes a speech sound, as /ʃa/ and /tʃa/, or a nonspeech sound, as in plucks and bows (see Cutting and Rosner, 1974, for experiments that make this comparison directly). Additional support for this view may be found in the work of Miller, Pastore, Wier, Kelly, and Dooling (1974), who found that noise-buzz sequences with varying relative onset times were also perceived categorically. The stimuli varied in a manner analogous to the voice-onset-time continuum; thus there appears to be a comparable mode of perception for such stimuli regardless of whether they are speech or nonspeech.

It may be possible to consolidate the above two views by finding a common conceptual relationship between the pairs "auditory-phonetic" and "intensity-rise time." By the definition of the term "phonetic" established above, it is clear that plucks and bows are not phonetic, since they cannot be articulated by the human vocal tract. However, the status of plucks and bows with respect to the second part of the definition, that the sounds can be recoded into higher-order linguistic units, is less clear. Certainly, they are not part of spoken language: but they do comprise lower level components in the "language" of music, which, like human spoken language, can be divided into hierarchical levels of organization (e.g., pitch, timbre, and harmony; see Nattiez, 1975, for a more complete discussion of this problem).

It is important to determine the extent to which plucks and bows are processed in terms of their specific acoustic characteristics, their status as nonspeech, or perhaps the role they play in a hierarchically organized system of sound. Various data have been used to support the view that it is the status of sounds as speech or nonspeech, rather than specific acoustic characteristics, that determines whether certain kinds of processing will occur. For example, when variations in acoustic dimensions analogous to those characteristic of certain phonemes--such as isolated second formants--are presented in a nonspeech context for identification and discrimination, perception is no longer categorical (Mattingly, Liberman, Syrdal, and Halwes, 1971). However, such sounds are not only nonlinguistic, they also bear little resemblance to sounds that commonly occur in the listener's environment. In contrast, the plucks and bows used in this experiment, which are not phonetic by the strictest definition, do resemble commonly heard musical sounds. It is an intriguing possibility that the plucks and bows are perceived in ways similar to speech partly because of their "codability" or "meaningfulness" for the listener. Perhaps it is the interaction of the acoustic nature of the sound and its significance for the listener that leads to the kinds of perceptual phenomena that have been considered exclusive to speech.

The results of experiments using noise-buzz stimuli (Miller et al., 1974) appear to argue against the analogy between the perception of basic musical units and phonemes. However, unlike plucks and bows, the noise-buzz sequences have been shown to be perceived as speech in only one experimental paradigm. If, in fact, they did match the speech results in other paradigms, such as selective adaptation and speeded classification, one would still want to ascertain whether subjects phenomenologically experience the stimuli as resembling common environmental sounds that are codable in a hierarchically organized system of sound like music and language.[4]

In conclusion, we do not question that levels of processing separate certain linguistic and nonlinguistic dimensions of the same stimuli. We suggest, rather, that the crux of the auditory-phonetic distinction is, as Fry (1956) suggested, a "physical-psychological transformation." The nature of this transformation probably cannot be accounted for solely in terms of the linguistic-nonlinguistic distinction. Instead it may be based on acoustic properties alone, on the coding of sounds within a hierarchically organized system, or on the interaction of acoustic properties with such a system.

REFERENCES

Biederman, I. and S. F. Checkosky. (1970) Processing redundant information. J. Exp. Psychol. 83, 486-490.

Chaney, R. B. and J. C. Webster. (1966) Information in certain multidimensional sounds. J. Acoust. Soc. Amer. 40, 447-455.

Chomsky, N. and M. Halle. (1968) The Sound Pattern of English. (New York: Harper & Row).

Cooper, F. S. and I. G. Mattingly. (1969) Computer controlled PCM system for investigation of dichotic speech perception. J. Acoust. Soc. Amer. 46, 115(A).

Cooper, W. E. (1975) Selective adaptation to speech. In Cognitive Theory, ed. by F. Restle, R. M. Shiffrin, N. J. Castellan, H. Lindman, and D. B. Pisoni. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.).

Curry, F. K. W. (1967) A comparison of left-handed and right-handed subjects on verbal and non-verbal dichotic listening tasks. Cortex 3, 343-352.

Cutting, J. E. (in press) The magical number two and the natural categories of speech and music. In Tutorial Essays in Psychology, vol. 1, ed. by N. S. Sutherland. (Hillsdale, N. J.: Lawrence Erlbaum Assoc.). [Also in Haskins Laboratories Status Report on Speech Research SR-42/43 (this issue).]

---

[4] In fact, several subjects in the present experiment and other experiments using plucks and bows have reported that they coded the sounds in very personal ways. For example, one subject found it helpful to think of the plucks as piano sounds and the bows as organ tones, even though such a distinction would normally be cued by more than just rise time.

Cutting, J. E. and B. S. Rosner. (1974) Categories and boundaries in speech and music. Percept. Psychophys. 16, 564-570.

Cutting, J. E., B. S. Rosner, and C. F. Foard. (1975) Rise time in nonlinguistic sounds and models of speech perception. Haskins Laboratories Status Report on Speech Research SR-41, 71-93.

Day, R. S. and C. C. Wood. (1972) Interactions between linguistic and nonlinguistic processing. J. Acoust. Soc. Amer. 51, 79(A).

Eimas, P. D., W. E. Cooper, and J. D. Corbit. (1973) Some properties of linguistic feature detectors. Percept. Psychophys. 13, 247-252.

Eimas, P. D. and J. D. Corbit. (1973) Selective adaptation of linguistic feature detectors. Cog. Psychol. 4, 99-109.

Fant, G. (1967) Auditory patterns of speech. In Models for the Perception of Speech and Visual Form, ed. by W. Wathen-Dunn. (Cambridge, Mass.: MIT Press).

Felfoldy, G. L. and W. R. Garner. (1971) The effects on speeded classification of implicit and explicit instructions regarding redundant dimensions. Percept. Psychophys. 9, 289-292.

Fry, D. B. (1956) Perception and recognition in speech. In For Roman Jakobson, ed. by M. Halle, H. G. Lunt, and C. H. van Schoonveld. (The Hague: Mouton).

Garner, W. R. (1974) The Processing of Information and Structure. (Potomac, Md.: Lawrence Erlbaum Assoc.).

Garner, W. R. and G. L. Felfoldy. (1970) Integrality of stimulus dimensions in various types of information processing. Cog. Psychol. 1, 225-241.

Garner, W. R., H. W. Hake, and C. W. Eriksen. (1956) Operationism and the concept of perception. Psychol. Rev. 63, 149-159.

Jakobson, R., C. G. M. Fant, and M. Halle. (1963) Preliminaries to Speech Analysis. (Cambridge, Mass.: MIT Press).

Kimura, D. (1961) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.

Kimura, D. (1964) Left-right differences in the perception of melodies. Quart. J. Exp. Psychol. 16, 355-358.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, A. M., P. C. Delattre, F. S. Cooper, and L. J. Gerstman. (1954) The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monogr. 68, 1-13.

Liberman, A. M., K. S. Harris, H. S. Hoffman, and B. C. Griffith. (1957) The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358-368.

Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: Acoustical measurements. Word 20, 384-422.

Mattingly, I. G., A. M. Liberman, A. K. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.

Miller, G. A. (1956) The magical number seven plus or minus two, or, some limits on our capacity for processing information. Psychol. Rev. 63, 81-97.

Miller, J. D., R. E. Pastore, C. C. Wier, W. J. Kelly, and R. J. Dooling. (1974) Discrimination and labeling of noise-buzz sequences with varying noise-lead times. J. Acoust. Soc. Amer. 55, 390(A).

Nattiez, J. J. (1975) Sémiologie musicale: L'état de la question. Acta Musicologica 42, 153-171.

Nickerson, R. S. (1971) Binary-classification reaction time: A review of some studies of human information-processing capabilities. Psychon. Monogr., Suppl. 4, 17, Whole No. 65.

Posner, M. I. (1964) Information reduction in the analysis of sequential tasks. *Psychol. Rev. 71*, 491-504.

Shankweiler, D. P. and M. Studdert-Kennedy. (1967) Identification of consonants and vowels presented to the left and right ears. *Quart. J. Exp. Psychol. 19*, 59-63.

Stevens, K. N. and M. Halle. (1967) Remarks on analysis by synthesis and distinctive features. In *Models for the Perception of Speech and Visual Form*, ed. by W. Wathen-Dunn. (Cambridge, Mass.: MIT Press).

Stevens, K. N. and A. S. House. (1972) Speech perception. In *Foundations of Modern Auditory Theory*, vol. 2, ed. by J. V. Tobias. (New York: Academic Press).

Studdert-Kennedy, M. (in press) Speech perception. In *Contemporary Issues in Experimental Phonetics*, ed. by N. J. Lass. (New York: Academic Press).

Studdert-Kennedy, M., D. P. Shankweiler, and D. B. Pisoni. (1972) Auditory and phonetic processes in speech perception: Evidence from a dichotic study. *Cog. Psychol. 3*, 455-466.

Townsend, J. T. (1971) A note on the identifiability of parallel and serial processes. *Percept. Psychophys. 10*, 161-163.

Turvey, M. (1973) On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. *Psychol. Rev. 80*, 1-52.

Wood, C. C. (1974) Parallel processing of auditory and phonetic information in speech perception. *Percept. Psychophys. 15*, 501-508.

Wood, C. C. (1975) Auditory and phonetic levels of processing in speech perception: Neurophysiological and information-processing analyses. *J. Exp. Psychol.: Human Perception and Performance 104* (now vol. 1), 3-20.

Wood, C. C. and R. S. Day. (1975) Failure of selective attention to phonetic segments in consonant-vowel syllables. *Percept. Psychophys. 17*, 346-350.

Predicting Initial Cluster Frequencies by Phonemic Difference

James E. Cutting*

ABSTRACT

The frequency of occurrence for stop-liquid and stop-semivowel clusters can be predicted on the basis of the number of distinctive features that separate the member phonemes: the greater the phonemic difference, the more frequent the cluster. Predictions made in this manner are generally much better than those made from chance co-occurrence of successive phonemes. Assessments are made on four extensive corpora. Each of six distinctive features is examined individually.

Why does the phonology of a given language permit certain phoneme clusters to occur, but not others? Why do some clusters occur frequently, others rarely? Saporta (1955), among others, suggested that the answer to both questions may lie in an application of Zipf's law (Zipf, 1949): "The relative frequency of consonant clusters will reveal a tendency on the part of any language system to produce speech in such a way as to consider the effort of both the speaker and the listener" (Saporta, 1955:25). Zipf (1935) had applied this principle to descriptions of phonemes, but Saporta (1955; Keller and Saporta, 1957) first applied it to phoneme clusters. Unique to Saporta's approach was the conjoining of a phonetic feature system (Jakobson, Fant, and Halle, 1951) and the principle of least effort. He suggested that phoneme cluster frequency correlated with the number of distinctive features not shared between the member phonemes of a given cluster. Altmann (1969) termed this measure phonemic difference. Saporta purposefully excluded clusters with liquids (/l/ and /r/) and semivowels (/w/ and /y/) and suggested that these combinations needed further study. The present paper is concerned with these clusters and with predicting their frequency of occurrence by the principle of phonemic difference.

Carroll (1958) criticized Saporta's analyses on the grounds that inadequate consideration was given to the possible chance nature of the results. After performing the proper analyses, Carroll concluded that phonemic difference has merit as a measure for predicting cluster frequency, but that a much more extensive data base (cluster count) should be used in further research. Following this advice, the present investigation uses data available from several extensive corpora gathered since the publication of the Saporta and Carroll articles. In addition, the principle of phonemic difference is matched against a control

---

*Also Wesleyan University, Middletown, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-42/43 (1975)]

(or chance-factor) principle, where cluster frequency is predicted from the frequency of occurrence of the member phonemes.

The present research, however, differs in several ways from previous studies. First, it is concerned with only a selected number of clusters where phonemic difference is not extensive. The first member of all clusters is a stop consonant /p,b,t,d,k,g/, and the second member is either a liquid or a semivowel /l,r,w,y/. This limitation removes from consideration the more widely differing clusters used by Saporta and Carroll, which may follow a different principle than those investigated here. Second, this limitation alters the main hypothesis. Saporta suggested that the most frequently occurring clusters should be those with intermediate phonemic difference. Here, I hope to demonstrate that for certain clusters maximal difference in the number of distinctive features shared by successive phonemes serves to predict cluster frequency. It should be remembered, however, that Saporta's intermediate differences and the maximal differences presented in this paper are nearly the same: interphoneme dissimilarities along five features. Third, the present interpretation of phonemic differences is not linked to the principle of least effort. Saporta (1955) noted that for the speaker the situation of least effort should be one in which successive phonemes are most similar, but for the listener that situation is one in which the phonemes are least similar. This led him to predict that intermediate phonemic difference should be important, serving both speaker and listener. However, Wang (1959) found that the least effort principle did not apply to the perceptions of final clusters, thus questioning the role of the listener in this application of Zipf's law. Least efforts for the speaker may be difficult to assess in an unbiased fashion. Therefore, it seems unwise to shackle a principle of phonemic difference to the broader principle of least effort; instead, it should stand alone. Fourth, the present study is concerned only with initial clusters rather than with both initial and final. The major reason for this limitation is simply that there are very few liquid-stop and semivowel-stop clusters in English.

Before presenting evidence supporting a revised phonemic difference principle, a few matters concerning methodological approach must be mentioned. First, the particular distinctive features system used is that proposed by Halle (1964), elaborated from earlier versions (Jakobson, Fant, and Halle, 1951; [illegible]herry, Halle, and Jakobson, 1953). The subsequent feature system of Chomsky and Halle (1968) is rejected on the grounds that many of the additional features are redundant for the particular phonemes selected here. The values for each stop, liquid, and semivowel, are made explicit by Wickelgren (1966) using Halle's definitions, and they are shown in Table 1. Second, each feature is considered equally important. Such an assumption is dangerous. Carrol (1958), for example, suggested that each feature should be considered separately, since voicing alone contributed extensively to Saporta's main finding. Separate analyses can confirm whether phonemic difference is a general principle for phoneme clusters, or whether it is limited only to certain contrasts. To demonstrate the generality of this principle, proper analyses will be performed. Third, the present paper considers all stop-liquid and stop-semivowel combinations to be true clusters. This assertion goes against the view that stop + /y/ combinations, for example, may not be legitimate clusters since they only occur before the vowel /u/ (Hofmann, 1967; Chomsky and Halle, 1968), or that /kl, kr, kw/ combinations might be more parsimoniously described as single consonants (Hofmann, 1967; Menyuk and Klatt, 1968; Menyuk, 1972). For a review of the phoneme-versus-cluster controversy, see Devine (1971). Fourth, in order to predict cluster frequencies and

TABLE 1: Distinctive feature representation from Halle (1964; see also Wickelgren, 1966) of certain phonemes in English that can form initial clusters.

	Stop consonants						Liquids		Semivowels	
	Labials		Alveolars		Velars					
	/p/	/b/	/t/	/d/	/k/	/g/	/l/	/r/	/w/	/y/
Vocalic	-	-	-	-	-	-	+	+	-	-
Consonantal	+	+	+	+	+	+	+	+	-	-
Grave	+	+	-	-	+	+	-	-	+	-
Diffuse	+	+	+	+	-	-	+	-	+	+
Voiced	-	+	-	+	-	+	+	+	+	+
Continuant	-	-	-	-	-	-	+	+	+	+

to confirm these estimates, one needs accurate assessments of actual cluster occurrence. Four extensive corpora were selected: Denes (1965), a corpus including 72,000 phonemes spoken in British English from "phonetic readers" used to teach English to foreign students; Hultzén, Allen, and Miron (1964), a corpus of 20,000 phonemes spoken in General American from selected material in eleven different dramatic plays; Roberts (1965), an analysis of a word list including 15 million entries phonemically transcribed in General American; and Trnka (1966), an analysis of a phoneme count from 100,000 words of connected material transcribed in British English. Fifth, to provide a prediction of cluster frequency based on chance co-occurrence of successive phonemes, accurate assessments of the frequency of each of the ten phonemes in question are needed. The four sources cited above provide these estimates. They agree fairly well with others (Hayden, 1950; Tobias, 1959; Delattre, 1966; Card and Eckler, 1975). See Gerber and Vertin (1969) and Wang and Crawford (1960) for comparisons.

Table 2 displays the phonemic difference between member phonemes for the 24 clusters, along with observed and predicted percentages. Cluster frequencies were determined separately for each corpus. Observed percentages were then calculated by dividing the number of occurrences of each cluster by the total occurrences of all 24 clusters. To compute predicted percentages, the percentage occurrence for each of the ten individual phonemes was first obtained; the product of the percentages for the two phonemes in each cluster was determined; this product was divided by the sum of the products for all 24 clusters, and multiplied by one hundred.

Notice the variation among the four corpora. In particular, variation is greater for observed frequencies than for predicted frequencies. For example, observed frequencies for /pr/, /kr/, and /tl/ differ by factors of 3:1 or greater, whereas predicted frequencies differ by much less.

Eight correlation coefficients were calculated, two for each corpus. First, predicted and observed cluster frequencies were correlated within the same corpus, then the phonemic difference scores were correlated with the obtained cluster frequencies. The results of these analyses are shown in Table 3, along with statistical assessments of each. In addition, the mean predicted and mean observed frequencies were calculated from the four corpora, and correlations

TABLE 2: Twenty-four initial clusters, their phonemic difference, and their observed and predicted frequencies from four different corpora.

	Phonemic difference	Denes (1965) observed	Predicted from Denes (1965)	Hultzén et al. (1964) observed	Predicted from Hultzén et al. (1964)	Roberts (1965) observed	Predicted from Roberts (1965)	Trnka (1966) observed	Predicted from Trnka (1966)
/pr/	5	11.1	2.3	13.3	1.9	25.7	3.0	9.0	5.6
/pl/	4	10.7	3.1	8.3	1.1	6.5	1.5	6.0	3.0
/pw/	3	.1	2.1	.0	2.1	.0	2.1	.0	1.7
/py/	4	.5	1.2	2.2	3.0	.7	3.1	.9	.5
/br/	4	4.6	2.7	6.6	2.4	7.8	3.1	10.4	5.0
/bl/	3	9.6	3.6	3.9	1.4	4.7	1.5	8.1	2.7
/bw/	2	.0	2.5	.0	2.6	.0	2.1	.0	1.5
/by/	3	.1	1.4	.5	3.8	.5	3.1	.4	.4
/tr/	4	11.9	10.8	11.6	10.0	12.4	13.1	11.6	19.8
/tl/	3	10.4	14.2	4.4	6.0	.0	6.5	.0	10.6
/tw/	4	1.5	10.0	1.1	10.8	1.9	9.0	2.7	6.0
/ty/	3	2.4	5.8	.0	16.0	.0	13.4	1.3	1.7
/dr/	3	2.8	5.4	6.6	4.1	6.6	5.7	5.9	12.0
/dl/	2	2.5	7.1	2.2	2.5	.0	2.9	.0	6.4
/dw/	3	.0	5.0	1.7	4.5	.2	3.9	.4	3.6
/dy/	2	4.0	2.9	.0	6.5	.0	5.9	1.0	1.1
/kr/	4	2.4	3.7	8.8	3.5	8.8	4.6	10.8	7.5
/kl/	5	7.1	4.9	6.1	2.1	7.3	2.3	9.2	4.1
/kw/	4	7.6	3.5	5.0	3.8	4.5	3.2	5.0	2.3
/ky/	5	2.4	2.0	2.2	5.6	.7	4.7	.9	4.1
/gr/	3	6.2	1.5	12.2	1.4	9.1	1.6	10.2	1.9
/gl/	4	1.4	2.0	1.7	.9	2.6	.8	5.6	1.1
/gw/	3	.1	1.4	.0	1.6	.0	1.1	.4	.6
/gy/	4	.5	.8	1.7	2.3	.0	1.7	.0	.2
Total		99.9	99.9	100.1	99.9	100.0	99.9	99.8	99.9

TABLE 3: Correlations between observed cluster frequencies and (a) those predicted from phoneme frequencies within the same corpus, and (b) number of distinctive features separating phonemes within a cluster. df = N-3 = 21 (see McNemar, 1969).

Corpus	Correlation with frequencies predicted by chance	Correlation with phonemic difference
Denes (1965)	.40	.34
Hultzén, Allen, and Miron (1964)	-.19	.45*
Roberts (1965)	.03	.52+
Trnka (1966)	.45*	.46*
Mean of corpora	.20	.47*

*p<.05, two-tailed
+p<.01, two-tailed

performed. Notice that for the American corpora the observed frequencies correlated more highly with the phonemic difference scores than did the control, or chance-factor, frequency estimates. (This is all the more impressive since, owing to the large number of ties in phonemic difference, maximum correlation coefficients calculated here can be only .90.) The phonemic difference correlations for the Hultzén, Allen, and Miron (1964) corpus and the Roberts (1965) corpus were significantly greater than the control correlations (t = 2.51, p<.025 and t = 2.33, p<.05, respectively; McNemar, 1969:157-158). It is interesting to note that these are the two corpora based on American English pronunciation, whereas the other two are based on British English.

Is this principle general and distributed across the various phonetic features? Or is it, as Carroll (1958) suggests, primarily a function of one feature?--voicing. The answer can be seen in Table 4. The percent occurrence of all clusters that do not share each of the six features is compared against chance, calculated by dividing the number of clusters involved by the total number of clusters (24). Phonemic difference along all but one feature, consonantal, fits into the general scheme, providing equal to or greater than chance prediction. Consider the features in more detail. Vocalic and consonantal features can be yoked since all clusters differ on one or other (but not both) of the features; the first separates /l,r/, the second /w,y/, from the stop consonants. The vocalic feature predicts cluster frequencies very nicely, but the consonantal feature does not. Notice that the observed average of these two features is exactly 50 percent, or chance. Following Carroll (1958), one can eliminate these two features, yoked together, since they do not provide greater-than-chance prediction. The continuant feature can also be dismissed since members of all 24 clusters differ along this dimension. Only three features remain: grave, diffuse, and voiced. Each of these three features appears to contribute nearly equally to the phonemic difference effect, providing from 11 to 18 percent better-than-chance prediction. Moreover, voicing is not the most potent feature, as Carroll suggested it might be.

TABLE 4: Analysis of the effect of phonemic difference for each of the six distinctive features in predicting cluster frequency. Data base is mean of four corpora.

Distinctive feature	Observed percent of clusters with this feature not shared		Chance	Clusters involved
Vocalic	86.3	100.0	50.0	stop + /l,r/
Consonantal	13.7		50.0	stop + /w,y/
Grave	69.9		58.3	/pl, pr, py, bl, br, by, tw, dw, kr, kl, ky, gr, gl, gy/
Diffuse	59.2		41.7	/pr, br, tr, dr, kl, kw, ky, gl, gw, gy/
Voiced	66.1		50.0	/p,t,k/ + liquid /p,t,k/ + semivowel
Continuant	100.0		100.0	all

If the phonemic-difference principle is tenable as a predictor of cluster frequency, one might expect that any phonological change within a cluster should be in the direction of increasing phonemic difference, or perhaps in the elimination of the cluster altogether. In American English an increase can be seen in the affrication of alveolar + /r/ clusters; /tr/ and /dr/ clusters tend to go to /tʃr/ and /dʒr/, as in TRY and DRY. Affrication, in effect, adds the additional contrast of strident-nonstrident to members of both clusters. According to the mean of the four corpora, these two clusters are the second and tenth most frequent of the 24, and the addition of the strident contrast increases their phonemic difference to 5 and 4, respectively. In American English the stops /t/ and /d/ are involved in cluster elimination as well. Often these are simplified to a single consonant: /ty/ and /dy/ go to /t/ and /d/ in TUBE and DUTY. Notice that the Hultzén et al. (1964) corpus and the Roberts (1965) corpus lack any /ty/ and /dy/ clusters, whereas the Denes (1965) and Trnka (1966) corpora contain a number of them, which illustrates one difference between American and British English.

In conclusion, a revised version of the phonemic-difference principle postulated by Saporta (1955) serves to predict cluster frequency. Those clusters investigated here are stop-liquid and stop-semivowel combinations, which Saporta did not consider. As Carroll (1958) suggested, this principle serves to predict cluster frequency better than chance, particularly for the American English corpora. The effect is distributed across the different distinctive features and is particularly strong for grave, diffuse, and voicing features.

REFERENCES

Altmann,G. (1969) Differences between phonemes. Phonetica 19, 118-132.

Card, L. E. and A. R. Eckler. (1975) A survey of letter-frequencies. Word Ways: Journal of Recreational Linguistics 8, 81-85.

Carroll, J. B. (1958) The assessment of phoneme cluster frequency. Language 34, 267-278.

Cherry, E. C., M. Halle, and R. Jakobson. (1953). Toward the logical description of languages in their phonemic aspect. Language 29, 34-46.

Chomsky, N. and M. Halle. (1968) The Sound Pattern of English. (New York: Harper & Row).

Delattre, P. (1966) Comparing the Phonetic Features of English, French, German, and Spanish. (Heidelberg: Julius Groos Verlag).

Denes, P. B. (1965) On the statistics of spoken English. J. Acoust. Soc. Amer. 35, 892-904.

Devine, A. M. (1971) Phoneme or cluster: A critical review. Phonetica 24, 65-85.

Gerber, S. E. and S. Vertin. (1969) Comparative frequency counts of English phonemes. Phonetica 19, 133-141.

Halle, M. (1964) On the bases of phonology. In The Structure of Language, ed. by J. A. Fodor and J. J. Katz. (Englewood Cliffs, N. J.: Prentice-Hall), pp. 324-333.

Hayden, R. E. (1950) The relative frequency of phonemes in general American English. Word 6, 217-223.

Hofmann, T. R. (1967) Initial clusters in English. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 84, 263-274.

Hultzén, L., J. Allen, and M. Miron. (1964) Tables of Transitional Frequencies of English Phonemes. (Urbana: University of Illinois Press).

Jakobson, R., C. G. M. Fant, and M. Halle. (1951) Preliminaries to Speech Analysis. (Cambridge, Mass.: MIT Press, 1963).

Keller, K. C. and S. Saporta. (1957) The frequency of consonant clusters in Chontal. Intl. J. Amer. Ling. 23, 28-38.

McNemar, Q. (1969) Psychological Statistics, 4th ed. (New York: John Wiley & Sons).

Menyuk, P. (1972) Clusters as single underlying consonants: Evidence from children's productions. In Proceedings of the Seventh International Congress of Phonetic Sciences, ed. by A. Rigault and R. Charbonneau. (The Hague: Mouton), pp. 1161-1165.

Menyuk, P. and D. Klatt. (1968) Child's production of initial consonant clusters. Quarterly Progress Report (Research Laboratory of Electronics, MIT) 81, 205-213.

Roberts, A. H. (1965) A Statistical Analysis of American English. (The Hague: Mouton).

Saporta, S. (1955) Frequency of consonant clusters. Language 31, 25-30.

Tobias, J. V. (1959) Relative occurrence of phonemes in American English. J. Acoust. Soc. Amer. 31, 631.

Trnka, B. (1966) A Phonological Analysis of Present-Day Standard English. (University: University of Alabama Press).

Wang, W. S-Y. (1959) Transition and release as perceptual cues for final plosives. J. Speech Hearing Res. 2, 66-73.

Wang, W. S-Y. and J. Crawford. (1960) Frequency studies of English consonants. Lang. Speech 3, 131-139.

Wickelgren, W. A. (1966) Distinctive features and errors in short-term memory for English consonants. J. Acoust. Soc. Amer. 39, 388-398.

Zipf, G. K. (1935) The Psycho-biology of Language. (Cambridge, Mass.: MIT Press, 1965).

Zipf, G. K. (1949) Human Behavior and the Principle of Least Effort. (Cambridge, Mass.: Addison-Wesley).

Hemispheric Specialization for Speech Perception in Four-Year-Old Children from Low and Middle Socioeconomic Classes*

Donna S. Geffner[+] and M. F. Dorman[++]

ABSTRACT

Four-year-old male and female children from low and middle socioeconomic classes (SEC) were tested on a dichotic syllable task. Both low and middle SEC males evidenced significant right-ear advantages. Neither low nor middle SEC females evidenced a significant right-ear advantage. The similar ear advantage in the low and middle SEC populations replicates a previous study with six-year-olds and suggests that variations in rearing conditions between low and middle socioeconomic classes do not affect hemispheric lateralization for speech perception. The absence of a right-ear advantage for females replicates the outcome of several other investigators and points to the need for longitudinal rather than cross-sectional studies of the development of cerebral lateralization.

INTRODUCTION

Several studies have indicated that children from low socioeconomic class (SEC) backgrounds may develop cerebral lateralization for speech perception at a slower rate than their middle SEC cohorts. Kimura (1967), using a dichotic digits task to assess cerebral lateralization, reported that at age five both low and high SEC females and high SEC males evidenced a right-ear advantage (indicating left-hemisphere specialization for speech perception). Low SEC males, however, did not evidence a right-ear advantage until age six. Geffner and Hochberg (1971), also using a dichotic digits task, reported a significant right-ear advantage for middle SEC children aged four through seven, but found for low SEC children a significant right-ear advantage only at age seven. Taken together, these data suggest that some, as yet unspecified, environmental rearing conditions may retard the onset of cerebral lateralization of function.

Recently Dorman and Geffner (1974) have provided another interpretation of the studies reported above: the reduced right-ear advantage of the low SEC children on the dichotic digits tasks may be due to their overall poor performance,

---

*To be published in Cortex.

[+] Herbert H. Lehman College of the City University of New York.

[++] Also Herbert H. Lehman College of the City University of New York.

[HASKINS LABORATORIES: Status Report on Speech Research SR-42/43 (1975)]

or, in other words, to a "floor effect" reflecting task difficulty and lack of motivation. To assess this hypothesis, six-year-old black and white children from low and middle SEC backgrounds were tested with a simplified dichotic listening task (one monosyllable pair on each trial) by an experimenter of the subjects' own race. All groups evidenced a significant right-ear advantage. Furthermore, the magnitude of the right-ear advantage did not differ as a function of race or SEC. This outcome suggests that low SEC children, at least by age six, do not lag behind middle SEC children in development of hemispheric lateralization for speech perception.

It is, of course, possible that rearing conditions may exert an influence on cerebral lateralization of function, but that such effects are detectable only in children younger than age six. To assess this, in the present study, four-year-old children from low and middle SEC backgrounds were tested on a dichotic syllable task.

METHOD

Subjects

The subjects were 44 four-year-old children: 21 low SEC (9 males and 12 females); and 23 middle SEC (11 males and 12 females). Socioeconomic class was determined by Hollingshead's Two Factor Index of Social Position (Hollingshead, 1957), which takes into account the parents' educational level and occupational status. All subjects were right-handed (handedness tasks are detailed in the Procedure) and had normal hearing with no known perceptual, neurological, speech, or language deficit. Children with bilingual background were not selected.

Apparatus

The speech signals were recorded and reproduced on a Panasonic RS279US stereo tape deck. The signals were presented to the children via matched and calibrated TDH-39 headphones. The output of each tape channel was monitored by a 1000-Hz calibration signal on each channel. Audiometric threshold tests were administered on a Maico MA-10 portable audiometer calibrated to International Standards Organization (ISO) measures.

Preparation of Stimuli

Six stop-consonant-vowel syllables /ba, da, ga, pa, ta, ka/ were generated on the Haskins Laboratories parallel-resonance speech synthesizer. Each stimulus was 300 msec in duration. Under computer control the six syllables were combined into their 15 possible pairs and recorded dichotically, in a fully balanced order, on magnetic tape. The resulting tape contained 60 syllable pairs with each member of a pair occurring twice on each channel. The interstimulus interval was four seconds.

Procedure

Each subject was tested individually in a quiet room. All subjects were first given an audiometric threshold test. Hearing level at 500, 1000, 2000, and 4000 Hz was assessed. If the hearing level between the two ears differed by 10 dB or more for two of the test frequencies, the subject was excluded from further testing. Handedness was determined by asking the subjects to perform three manual motor tasks: throwing a ball, cutting with scissors, and drawing a circle. Any subject who did not perform all three tasks with his right hand was not tested further.

After the preliminary examination, the subjects were presented binaurally with three repetitions of the syllables /ba, da, ga, pa, ta, ka/. The subjects were instructed to listen with both ears and report the syllable heard. Any subject unable to repeat the six syllables after the third repetition of the list was excluded from further testing. The subjects were then instructed to listen again with both ears and report the syllable they heard. (Since the subjects were not told that there were two different stimuli on these and the following dichotic trials, only one response was elicited.) The subjects were told that these sounds would sound "funny," but to continue reporting them as before. The subjects were then presented the 60-item test sequence, followed by a brief rest, then the 60-item test again. To control for possible channel effects, the headphones were reversed after the first 60-item test.

RESULTS

A subject's results were excluded from the data analyses if he/she did not complete the 120 trials or if he/she gave perseverative responses (the same syllable on most trials). On these criteria, the test results of 48 percent of the low SEC subjects were excluded from the data analyses. In contrast, none of the data from the middle SEC subjects was excluded.

Each subject's performance was scored in terms of the metric (R-L/R+L) × 100, where R is the total number of syllables correctly recalled from the right ear and L is the total number of syllables correctly recalled from the left ear. The mean scores for the two socioeconomic class groups, as a function of sex, are shown in Table 1.

TABLE 1: Average magnitude of the right-ear advantage in terms of (R-L/R+L) × 100 for male and female subjects from low and middle SEC groups.

	Low SEC	Middle SEC
Male	12.53	12.00
Female	0.60	-0.83

The magnitude of the right-ear advantage did not differ significantly between the low and middle SEC groups ($z = 1.02$, $\underline{p} > .05$). However, an overall sex effect was observed ($z = 2.07$, $\underline{p} < .02$), with males evidencing a significantly larger right-ear advantage than females.

The mean number of syllables correctly reported from each ear for the male and female subjects in the low and middle SEC groups is shown in Table 2. Both low and middle SEC male subjects evidenced a significant right-ear advantage. Neither the low nor middle SEC female subjects evidenced a significant right-ear advantage. Sixty-six percent of the female subjects, collapsed over SEC, evidenced a right-ear advantage (mean = 11.22) and 33 percent a left-ear advantage (mean = 23.51). Of the male subjects, 65 percent evidenced a right-ear advantage (mean 22.04) and 35 percent a left-ear advantage (mean = 6.41). Thus, the male subjects evidenced both larger average right-ear advantages and smaller left-ear advantages than the female subjects.

TABLE 2: Mean number of syllables correctly reported from each ear.

Group	N	Left	Right	t
Male Middle SEC	11	32.82	41.73	2.05*
Female Middle SEC	12	41.08	38.75	-0.32
Male Low SEC	9	29.44	39.33	1.94*
Female Low SEC	12	35.58	36.41	0.20

*$p<.05$; one-tailed.

DISCUSSION

The presence of a similar right-ear advantage in both low and middle SEC four-year-old children replicates the outcome of an earlier study with six-year-olds (Dorman and Geffner, 1974). Thus, variation in rearing conditions, at least for the range subsumed under the categories low and middle SEC, does not appear to affect the rate of cerebral lateralization for speech perception (cf. Ingram, 1975). This conclusion must, however, be tempered by the fact that a large number of low SEC children could not be tested with the dichotic syllable task.

The absence of a significant right-ear advantage in females was unexpected. However, several other investigators, in spite of their different dichotic listening tasks, have reported a similar outcome with four-year-old females. Ingram (1975), Nagafuchi (1970), and Yeni-Komshian[1] have all found significant right-ear advantages for four-year-old males, but not for females. While one such effect may reasonably be attributed to sampling error, the similar outcome of four independent studies strongly suggests that females, aged four years, do indeed perform differently than male coevals on dichotic listening tasks.

One possible reason for this male-female difference, namely, that more females than males give left-ear advantages, is ruled out by the present study, since the proportion of subjects displaying a left-ear advantage was roughly the same for both sexes. Rather, an average right-ear advantage for females was eliminated because, while a majority of the children evidenced moderate right-ear advantages, the remainder displayed large left-ear advantages. Thus, the absence of an average right-ear advantage for the females does not imply that individual females are not lateralized. On the contrary, the absolute ear advantage (ignoring direction) for the males and females was essentially identical (females = 15.54; males = 16.67).

---

[1] 1973: personal communication.

Ingram (1975) has pointed out that the absence of an overall right-ear advantage in four-year-old females is all the more puzzling, given that three- and five-year-old females do display a right-ear advantage. Ingram noted one possible interpretation of these results--that there may be a period of cerebral reorganization, during which left-hemisphere functions are temporarily preempted by functions other than speech. Other interpretations are, of course, possible. For example, changes in magnitude of the ear-advantage may reflect changing linguistic processing strategies, similar to those demonstrated by Bever (1970) in sentence processing tasks. In any event, cross-sectional developmental studies, with their inherent problems of sampling error, are clearly inadequate to the task: a longitudinal cohort study (cf. Schaie and Strother, 1968) seems to be needed to resolve these issues.

REFERENCES

Bever, T. G. (1970) The cognitive basis for linguistic structures. In Cognition and the Development of Language, ed. by R. Hayes. (New York: Wiley).

Dorman, M. F. and D. S. Geffner. (1974) Hemispheric specialization for speech perception in six-year-old black and white children from low and middle socioeconomic classes. Cortex 10, 171-176.

Geffner, D. S. and I. Hochberg. (1971) Ear laterality performance of children from low and middle socioeconomic levels on a verbal dichotic listening task. Cortex 7, 193-203.

Hollingshead, A. (1957) Two Factor Index of Social Position. (New Haven, Conn.).

Ingram, D. (1975) Cerebral speech lateralization in young children. Neuropsychologia 13, 103-105.

Kimura, D. (1967) Functional asymmetry of the brain in dichotic listening. Cortex 3, 163-178.

Nagafuchi, M. (1970) Development of dichotic and monaural hearing abilities in young children. Acta Otolaryngol. 69, 409-414.

Schaie, K. W. and C. R. Strother. (1968) A cross-sequential study of age change in cognitive behavior. Psychol. Bull. 70, 671-680.

Automatic Segmentation of Speech into Syllabic Units*

Paul Mermelstein

ABSTRACT

As a first step toward automatic phonetic analysis of speech, one desires to segment the signal into syllable-sized units. Experiments were conducted in automatic segmentation techniques for continuous, reading-rate speech to derive such units. A new segmentation algorithm is described that allows assessment of the significance of a loudness-minimum to be a potential syllabic boundary from the difference between the convex hull of the loudness function and the loudness function itself. Tested on roughly 400 syllables of continuous text, the algorithm results in 6.9 percent syllables missed and 2.6 percent extra syllables relative to a nominal, slow-speech syllable count. It is suggested that inclusion of alternative fluent-form syllabifications for multisyllabic words and the use of phonological rules for predicting syllabic contractions can further improve agreement between predicted and experimental syllable counts.

INTRODUCTION

Automatic phonetic analysis of speech, such as that carried out as part of a continuous speech understanding system, requires a mapping from acoustic signal to phonetic segments whose direct implementation has eluded speech researchers for many years. Liberman (1970) reviews the case for considering the conversion between phone and sound to be a process of complex grammatical recoding that may prevent one from ever finding a direct replacement of sound segments by phones. In agreement with that point of view, we consider an alternative, indirect approach that segments the speech stream into syllable-sized units and decodes the phonetic segments of those units by considering the acoustic information contained in the entire syllable (Mermelstein, 1975). This paper presents results of experiments in automatic segmentation of continuous speech into such syllable-sized units.

---

*Published in The Journal of the Acoustical Society of America (1975), 58, 880-883.

Acknowledgment: I wish to acknowledge with thanks discussions of the material presented here with my colleagues at Haskins Laboratories: F. S. Cooper, J. Gaitenby, G. Kuhn, and P. Nye. This research was supported in part by the Advanced Projects Agency of the Department of Defense under contract No. N00014-67-A-029-002 monitored by the Office of Naval Research. The views presented here do not necessarily represent the views of the Department of Defense.

The syllable has been defined linguistically as "a sequence of speech sounds having a maximum or peak of inherent sonority (that is, apart from factors such as stress and voice pitch) between two minima of sonority" (Robins, 1966). To arrive at an operational definition that can be implemented computationally, one must define sonority in terms of physical measures on the speech signal. This requirement leads quickly to the realization that "inherent sonority" cannot be empirically defined because the same parameter--intensity--signals (in part) both sonority and stress, and the division between the two factors is rather arbitrary. The argument that stress values are assigned to entire syllables and sonority varies from phone to phone within the syllable cannot be applied to separate the two factors since it is precisely the operational determination of syllables that we are trying to achieve.

Stowe (1963) attacked this problem by a hierarchic series of segmentation procedures, each operating on a different time function computed from the speech signal. Sargent, Li, and Fu (1974) also used two functions for syllable detection, one measuring peak-to-peak amplitude, the other root-mean-square (RMS) intensity. In this work we explore a new approach. We attack the resolution of the above problem by defining a "loudness" measure for the speech signal, a time-smoothed and frequency-weighted summation of its energy content. Relative loudness maxima are interpreted as potential syllabic peaks and relative loudness minima as potential syllabic boundaries. To differentiate between syllables generally defined on the phonological level and the speech segments that may be located in the signal by phonetic criteria, we introduce the term "syllabic unit" for the syllable-sized speech segments that are to be found automatically. Boundaries located by loudness criteria do not necessarily segment the speech signal at points that can be identified as phone boundaries, or even word boundaries. The syllabic units are found to depend strongly on the phonetic performance of the speaker; in fact, they serve to describe that performance by grouping segments into larger units that generally form units of production as well.

In order to arrive at a segmentation of the signal into syllable-sized units, we find that one must define a measure of significance that permits classifying loudness minima as to whether they denote actual boundaries. Otherwise, the number of realized segments greatly exceeds the number of syllables one would count perceptually. Further, the measure of significance must be a function of the context of any particular loudness minimum. A local loudness minimum separated by less than 100 msec from another local minimum with lesser loudness may be insignificant, yet the same minimum with no other minima within 500 msec would generally signal a syllabic boundary.

The significance of loudness maxima must be similarly evaluated. In order to prevent segmentation into fragments that do not contain adequately strong syllabic peaks, we reject any segment whose loudness maximum is more than a given threshold below the overall loudness maximum, the syllabic peak of the loudest syllable of the utterance. Similarly, a minimum syllabic-unit duration of 80 msec is imposed, and segmentation that would result in shorter fragments is rejected.

One important application of syllabic-unit segmentation is as an aid to lexical analysis where one would like the same text spoken by different speakers to show at most a small number of alternative syllabic-unit representations. Fricatives are generally not tightly bound to the syllabic units with which they are associated but are frequently separated from them by a short interval of weak voicing or even silence. On the basis of loudness criteria alone, they form valid syllabic units. For the purposes of evaluating the results of our segmentation procedures and for accessing a lexicon of syllabic forms, we require that syllabic units have nonfricative nuclei. If subsequent analysis reveals that a syllabic unit manifests significant frication near the syllabic peak, it is labeled as a syllabic fragment and not counted as an independent syllabic unit.

SEGMENTATION USING A CONVEX-HULL ALGORITHM

In order that our empirically determined loudness function roughly approximate the subjective loudness function, loudness is obtained from the speech power spectrum by weighting frequencies below 500 Hz and above 4 kHz according to a function that drops off at 12 dB/octave outside these frequencies. To eliminate variations in loudness due to the phase of the fundamental frequency of excitation, the loudness function is low-pass filtered at 40 Hz. Our implementation computes loudness from the short-time power spectrum, but it could be equally well derived by directly filtering the speech wave.

Initially, a segment of speech between apparent pauses (silent interval exceeds 200 msec) is selected for analysis. The convex hull of the loudness function is defined as the minimal-magnitude function that is monotonically nondecreasing from the start of the segment to its point of maximum loudness, and is monotonically nonincreasing thereafter. Within the segment, the difference between the convex hull and the loudness function serves as a measure of significance of loudness minima. The point of maximal difference is a potential boundary. If the difference there exceeds a given threshold, the segment is divided into two subsegments.

Segmentation is carried out recursively. The convex hulls newly computed for the subsegments nowhere exceed the convex hull of the original segment. Hence, after any segmentation step, only less significant minima remain. If the maximal hull-loudness difference within the segment is below the threshold, no further segmentation of that segment is attempted. The algorithm makes use of the loudness context implicitly by extracting minima in order of significance. Thus, a minimum may not be significant if there is a more significant one close by. Segmentation removes the more significant minimum and allows reconsideration of the significance of the other minimum.

Figure 1 illustrates how the implementation of the convex-hull algorithm is applied. An original speech segment over the interval (a-c) is found to possess a loudness function $\ell(t)$ with a maximum at point b. The convex-hull computed for the segment (a-b-c) is $h_1(t)$. Over the interval (a-c), the maximum hull-loudness difference is d_1 at c'. If d_1 exceeds the threshold, segment (a-b-c) is cut up into segment (a-c') followed by segment (c'-b-c). The hull for segment (a-c'), defined around the new maximum point b', follows the loudness curve. This results in a zero hull-loudness difference over that interval and that portion is not segmented further. The hull for segment (c'-b-c), denoted by $h_2(t)$, is shown by the short dashed line where it differs from $h_1(t)$ over the segment interval. The new maximum hull-loudness difference is found to be d_2. If d_2 does not exceed the threshold, then the segment (c'-c) is not divided further.

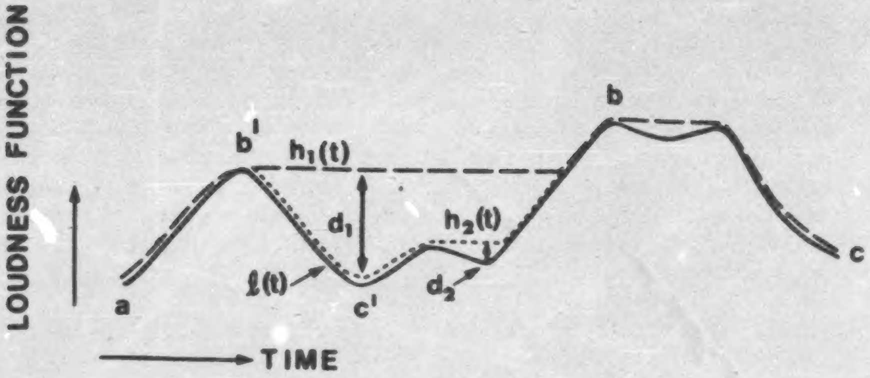
LOUDNESS FUNCTION
TIME
a
b'
h₁(t)
d₁
ℓ(t)
c'
h₂(t)
d₂
b
c

Figure 1: Loudness function and convex hull for a typical speech segment.

The algorithm does not proceed from left to right in time. It assumes that the entire utterance is stored before processing commences, but requires only that a complete segment delimited by silent intervals be captured before segmentation starts. Where real-time operation is essential, the algorithm can be modified to operate from left to right with possible backtracking over an interval of no larger than the maximum syllabic unit interval, roughly 500 msec.

Experimental Results

The performance of the algorithm was evaluated by processing 11 sentences read by each of two male subjects at their comfortable reading rate. The first six sentences (text A) make up the well-known "Rainbow Passage," and contain both monosyllabic and multisyllabic words. The last 5 (text B) consisted of only monosyllabic words and were taken from material composed by Lea (1974). The differentiation in text material was utilized to explore the dependence of segmentation errors on the frequency of multisyllabic words in the text.

Figure 2 illustrates typical results for the text "...a boiling pot of gold at...." The segmented loudness function is plotted above a computer-generated spectrographic representation of the utterance. The spectral data have been preemphasized at 6 dB/octave above 300 Hz. Use of a uniformly weighted intensity for the loudness function would miss the high-frequency energy discontinuity for [boj - liŋ]. By using loudness as defined, high-frequency energy variations are emphasized and the boundary is located.

By varying the segmentation threshold parameter d, we can control the relative frequency of extra syllabic units found and the frequency of syllables missed. A threshold d = 0 will result in too many extra syllabic units due to segmentation even at points of minimal variation in the speech loudness. A high threshold, d > 3 dB, will result in many significant segmentation points within voiced segments being missed. The segmentation results at d values of 2 dB, as compared to 1 dB, showed that 12 extra syllables in the corpus of 418 syllables had been eliminated, and only two new missed syllables had been introduced. Further small increases in the value of d did not result in any appreciable difference in performance; therefore, all further results are given for the d = 2 dB condition.

Differences between the output of the segmentation algorithm and a nominal syllable count are given in Table 1, classified by category and speaker. Since the syllable count is dependent on fricative detection, errors resulting from incorrect fricative detection are indicated separately, denoted by categories E2 and M2, respectively. The major source of extra syllabic units was in prepausal position (category E1) where significant release gestures were associated with final stops and liquids. The syllabic-unit loudness peaks for these cases were well above the -25-dB syllabic peak threshold, a value arrived at by empirical adjustments to eliminate most syllabic fragments. The frequency of prepausal extra syllabic units was highly speaker dependent, 1.2 percent for subject LL, only 0.5 percent for GK.

Syllabic units were missed primarily because of the tendency of an unstressed and stressed syllable-pair to contract into one stressed syllabic unit (category M1). Most such junctures had a loudness minimum that was not less than 1 dB below the last convex hull computed, some in fact had no loudness minima associated with them at all. In the monosyllabic text B, such errors

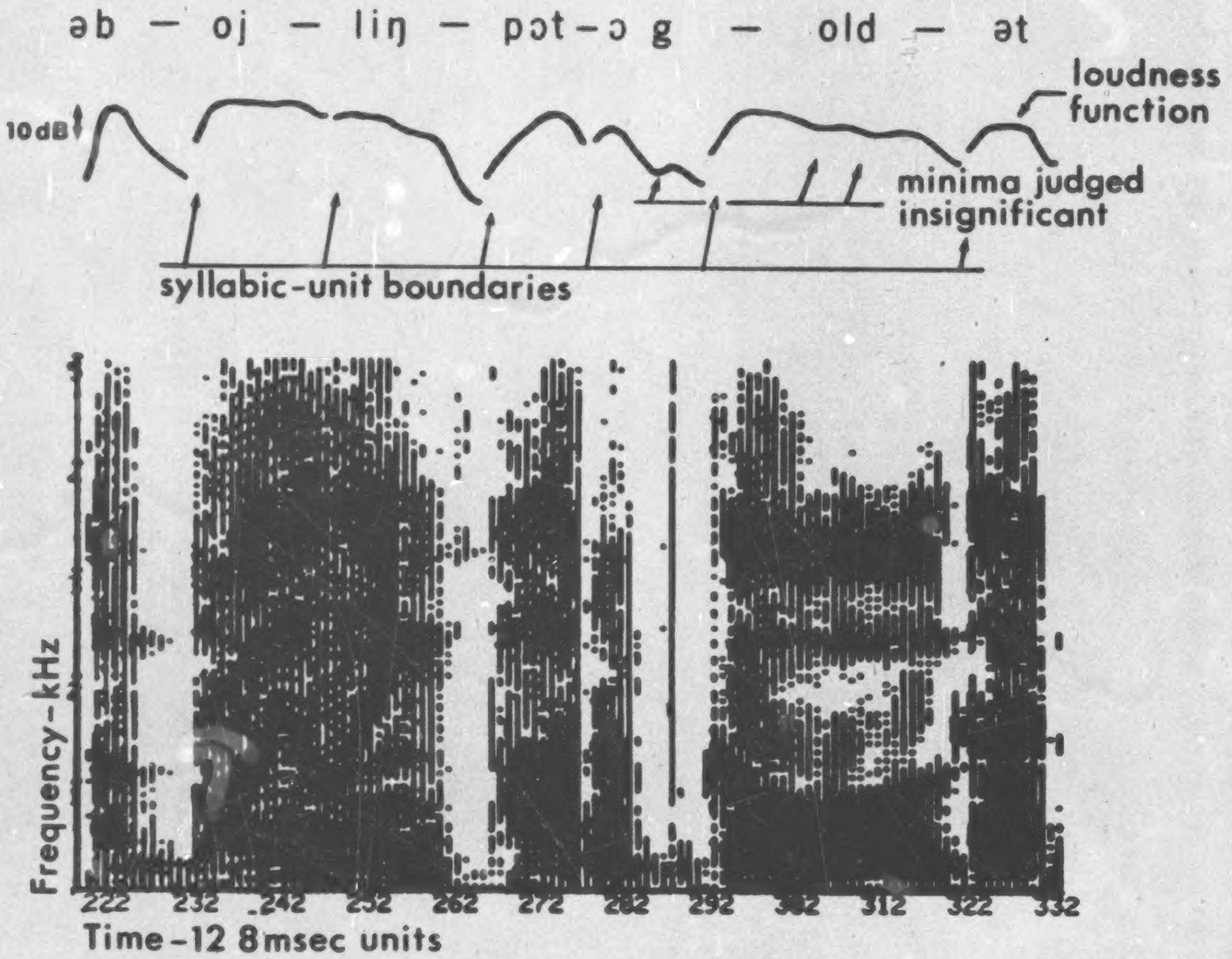

Figure 2: Example of syllabic segmentation results for text "...a boiling pot of gold at...."

L

TABLE 1: Differences between algorithm-derived syllable count and manual slow-speech count.

Speaker		Difference category									
		E1		E2		M1		M2		Total Syll.	
	Text	A	B	A	B	A	B	A	B	A	B
LL		1	1	1	-	12	2	2	2	123	86
GK		1	4	2	1	5	3	2	1	123	86

were encountered mostly in open syllables, e.g., /so aj/, /hɪ hᵊd/, but their frequency was rather low (0.6 percent). Possible contractions across words that may result in a syllable count for a sequence of words that is smaller than the sum of the individual syllable counts may best be handled on the phonological level through a set of rules predicting such phenomena. In the multisyllabic text the frequency of syllables missed was significantly higher (1.4 percent). Many of these (10 of 22) were encountered in both speakers' productions; that is, the syllable count for the same word or words for both speakers was consistent but different from a nominal syllable count that one would expect in slow speech. Typical examples are [hraj-zən] and [əp-pɛrn-lɪ] for "horizon" and "apparently." These forms must be considered to constitute acceptable productions alternative to those that would contain an additional syllabic unit for each word. Our results suggest a frequency of occurrence of problematic words whose syllabification cannot be adequately treated on the phonetic level. For speech recognition applications, it seems advisable to handle these multisyllabic word problems on the lexical level by including alternative admissible syllabifications in any lexicon of syllabic forms.

Differences categorized E2 and M2 denote extra and missed syllabic units due to incorrect categorization of the unit as nonfricative and fricative, respectively. Missed units result if a short vowellike interval is missed and the unit is interpreted as completely fricative, e.g., [tʰʊ], due to a previously discussed decision not to count fricativelike syllabic units as independent. Extra units result if a voiced fricative shows voicing sufficiently strong that it is interpreted as vowellike. Presumably these errors could be eliminated through improvements in the fricative detection algorithm.

In summary, the overall frequency of syllable-count differences with respect to nominal, slow-speech syllable count was 9.5 percent, consisting of 6.9 percent missed and 2.6 percent extra. We have previously reported 2.7 percent boundaries missed and 9 percent extra syllabic units found by essentially the same algorithm for a different text of 430 syllables (Mermelstein and Kuhn, 1974). There the algorithm was not optimized for the value of threshold d and errors were counted relative to a perceptual syllable count on the same spoken material. The difference in missed boundaries arises from the difference in the standard of comparison. For lexical applications a maximal syllabic form appears as the most useful standard.

The two sets of results, those reported here and those previously reported (Mermelstein and Kuhn, 1974), carried out on data from different speakers and collected under different recording conditions, yield roughly comparable difference rates. The previous study used data on a total of 31 sentences recorded by some five speakers, three male and two female. No large differences in overall syllable-count difference rates are observed as long as the speech is spoken carefully and at a moderate rate.

LOCATION OF SYLLABIC-UNIT BOUNDARIES

The boundaries located by the algorithm do not bear a simple relationship to general syllabification rules followed in "phonological" syllabic boundary assignment where the main criterion appears to be whether words occur in the language with that particular initial or final cluster. Based on these criteria, the syllabification of words containing intervocalic nasals and liquids is generally ambiguous, the sonorant may be assigned to either syllable-initial or syllable-final position. Linguists generally assign the maximum initial consonant sequence to the stressed syllable (Hoard, 1971). The algorithm locates a boundary within the consonant roughly at the point of minimal first-formant frequency. The major part of the consonantal segment is generally found assigned to the syllable carrying heavier stress due to its greater loudness. Where allophonic variations are associated with syllabic position, e.g., [ɹ̥'ɑʊnd] vs. [ə'rɑʊnd], the syllabification resulting from use of the algorithm is generally consistent with our phonetic expectations.

The change in loudness or intensity at the onset of a syllable is generally more abrupt than at its end; thus there is less uncertainty about the onset-time of a syllable than about its termination. Therefore, silent segments or those whose loudness is below the noise threshold are arbitrarily assigned to syllable-final position. This results in inclusion of nonreleased final stops in the previous syllable, but released stops straddle the syllabic boundary.

Intervocalic clusters are generally divided up. Compounds such as /sʌnlajt/ are segmented in accordance with morphemic criteria as loudness is found to decrease over the nasal and to increase over the liquid. Initial or final clusters may however be frequently broken up by the syllabification when unstressed syllables precede or follow them. For example, /tʊ/+/grit/ may map to [tʊg - rit], or /pajlz/+/ɔf/ to [pajl - zəf], where the symbol - indicates the position of the boundary within the phonetic segment stream. Generally the effect is to couple an initial or final cluster constituent with the preceding or following syllable if that ends or starts with a vowel. These effects occur sufficiently consistently, at least in our limited data, so that syllable reorganization may be predictable by rules.

The algorithm provides a useful tool for phonetic analysis. The word-pair /rezd/ vs. /redz/ forms an interesting example where attempts to use phonological criteria such as the measure of "vowel affinity" proposed by Fujimura (1975) to constrain the admissible syllable structures in English break down. Here /z/ and /d/ are phonemes that may occur in either order in syllable-final position, an exception to a general ordering of phonemes by increasing "vowel affinity" in syllable-initial and decreasing "vowel affinity" in syllable-final position. The convex-hull algorithm invariably classifies /rezd/ as one syllabic unit and /redz/ as two, a proper unit followed by a syllabic fragment. The "vowel affinity" of the fricative is different in the two cases, as manifested by a difference in intensity of voicing. The fricative in /redz/ is but weakly voiced, frequently devoiced. The postvocalic /z/ preceding a voiced stop carries stronger voicing. When followed by an unstressed vowel in syllable-initial position, this difference manifests itself by an assignment of the fricative to the first syllabic unit in the case where /rezd/+/ɪn/ gives [rezd - din] (syllabic-unit boundary within the closure of the /d/), but to the second syllabic unit in /redz/+/ɪn/ mapping to [red - zɪn]. We conclude that for the purposes of phonetic analysis, information derived regarding syllabic units and fragments is in fact useful even though for syllable-counting purposes one may desire to minimize the number of such fragments.

Conclusions

Syllabic units can be counted in continuous speech by simple automatic techniques. The number of syllabic units found will agree relatively reliably with a text-derived syllable count under the following conditions:

1. The algorithm is tuned to minimize extra syllabic units and missed units by adjusting the significance-threshold d.

2. A moderate amount of postprocessing is performed to weed out fricativelike syllabic fragments because they do not constitute independent syllabic units.

3. Phonological rules are employed to predict where separate words may be contracted to reduce the syllabic count of the total to less than the sum of the individual counts.

4. Alternative fluent-production forms are recognized for many multisyllabic words.

Segmentation into syllabic units appears to be sufficiently consistent so that the units so delimited constitute appropriate units of the speech signal on which further analyses may be carried out to extract additional phonetic information.

REFERENCES

Fujimura, O. (1975) Syllable as a unit of speech recognition. IEEE Trans. Acoust. Speech Sig. Proc. ASSP-23, 79-82.

Hoard, J. E. (1971) Aspiration, tenseness, and syllabication in English. Language 47, 133-139.

Lea, W. A. (1974) Sentences for controlled testing of acoustic phonetic components of speech understanding systems. Report PX 10952, Sperry Univac Defense Systems, St. Paul, Minn.

Liberman, A. M. (1970) The grammars of speech and language. Cog. Psychol. 1, 301-323.

Mermelstein, P. (1975) A phonetic-context controlled strategy for segmentation and phonetic labeling of speech. IEEE Trans. Acoust. Speech Sig. Proc. ASSP-23, 79-82.

Mermelstein, P. and G. M. Kuhn. (1974) Segmentation of speech into syllabic units. J. Acoust. Soc. Amer., Suppl. 55, 22(A).

Robins, R. H. (1966) General Linguistics: An Introductory Survey. (Bloomington: Indiana University Press).

Sargent, D. C., K. P. Li, and K. S. Fu. (1974) Syllable detection in continuous speech. J. Acoust. Soc. Amer. 55, 410(A).

Stowe, A. N. (1963) Segmentation of speech into syllables. J. Acoust. Soc. Amer. 35, 806(A).

On Pushing the Voice-Onset-Time (VOT) Boundary About*

Leigh Lisker,[+] Alvin M. Liberman,[++] Donna Erickson,[+++] and David Dechovitz[+++]

ABSTRACT

There is voluminous evidence that homorganic stop consonants are distinguishable on the basis of voice onset time (VOT) relative to their supraglottal articulation. For initial stops, a convenient acoustic reference point is the onset of the release burst, and VOT has been defined as the interval between this point and onset of glottal signal. Voice-onset-time boundary values between voiced and voiceless initial stops of English have been established by spectrographic measurements of naturally produced isolated words and by perception testing of synthesized consonant-vowel (CV) syllables. The close match between the two kinds of boundary values suggests that fairly natural values were chosen for the invariant features of the synthetic speech patterns tested. It is known, however, that certain of these affect voicing perception. New data from synthesis experiments show that VOT boundaries shift with changes in transition duration, that the first formant and not higher ones are responsible, and that transition duration is constrained to values that differ for place of articulation.

There is a good deal of evidence that homorganic stops are distinguishable on the basis of voice onset time (VOT) relative to their supraglottal articulation. For initial stops a convenient acoustic reference point is the onset of the release burst, and VOT has been defined as the interval between this point and the onset of glottal signal. Voice-onset-time boundary values between English initial /b,d,g/ and /p,t,k/ have been determined by spectrographic measurements of naturally produced isolated words and by perception testing of synthesized CV syllables. The close match between the two kinds of boundary values suggests that fairly natural values were chosen for features of the synthetic speech patterns that were not subjected to experimental manipulation in the

---

*Paper presented at the 89th meeting of the Acoustical Society of America, Austin, Tex., 7-11 April 1975.

[+] Also University of Pennsylvania, Philadelphia.

[++] Also University of Connecticut, Storrs, and Yale University, New Haven, Conn.

[+++] Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-42/43 (1975)]

perceptual experiments. It has been known for some time, however, that certain of these features do affect the perception of stop voicing. Thus, Cooper, Delattre, Liberman, Borst, and Gerstman (1952) pointed to a rising first formant as a cue to voicing, and the same group (Liberman, Delattre, and Cooper, 1958) later singled out first-formant "cutback" as a formidable cue to the English voiceless stops. More recently, Fujimura (1971) and Haggard, Ambler, and Callow (1970) have shown that fundamental frequency can also serve as a cue to the stop-voicing contrast. Most recently, Stevens and Klatt (1974) have emphasized the role of transition duration, showing that with greater durations there is an increase in the VOT value at the boundary between synthetic /da/ and /ta/ syllables. From all these studies, it is clear that listeners do not attend exclusively to VOT in judging synthetic stop-vowel patterns. (Here "VOT" signifies the duration of the interval between burst onset and the time when the periodic signal source is switched on and the first formant simultaneously shifts from zero to full amplitude.) Stevens and Klatt (1974) have argued that listeners, at least a significant proportion of them, respond in categorical manner to the presence versus absence of rapid frequency shifts in the formants, particularly the first formant following the onset of voicing. This not only accounts for their data, but also, as they point out, serves to explain why VOT boundaries vary with place of stop closure, since it has been observed that burst and transition durations also vary with place in natural speech. The Stevens-Klatt theory emphasizes the fact that there is another temporal landmark, aside from burst and voicing onset, that may have perceptual importance for stop voicing; namely, the point where formant frequencies achieve values appropriate to the following vowel. It might be the case, they seem to be saying, that the choice of the burst as the reference point for measuring VOT is more a matter of visual convenience for the spectrogram reader than of selecting the most useful landmark for the human auditor. At least two questions may be raised here: (1) Is the Stevens-Klatt hypothesis the only one suggested by their data? (2) Is their proposed new measure of VOT more nearly sufficient than VOT as a basis for categorizing the stops?

To help answer these questions, we can examine some new data that show, to begin with, that the Stevens-Klatt (1974) finding is in fact replicable. In Figure 1 we have the responses of seven phonetically naive young talkers of English asked to label as either /da/ or /ta/ a set of appropriately designed synthetic speech patterns. The variables are VOT and transition duration. Voice onset time was varied in 10-msec steps from 5- to 65-msec delay in onset of pulsing and first formant relative to the burst. In Test I six transitions, ranging from 20 to 85 msec in duration, were used; in Test II the durations ranged from 40 to 115 msec, this last value being the greatest for which acceptable /da/ and /ta/ syllables could be heard. Just as Stevens and Klatt found, the 50 percent crossover points along the VOT dimension move to higher values with increasing transition duration. The crossover for the shortest transition tested differs from the one for the longest by slightly more than 25 msec. This shift is just about twice as great as that reported by Stevens and Klatt. Of course the 25-msec shift shown here is occasioned, as we see, by a change of 95 msec in transition duration; in the Stevens-Klatt experiment the transitions were varied by only 30 msec. Insofar as they are comparable, our data show the closest possible agreement with theirs. However, on the basis of our data, it is as easy to emphasize the stability of the VOT boundary in the face of an extreme change in transition duration as it is to point out the undeniable fact that it is not absolutely immutable.

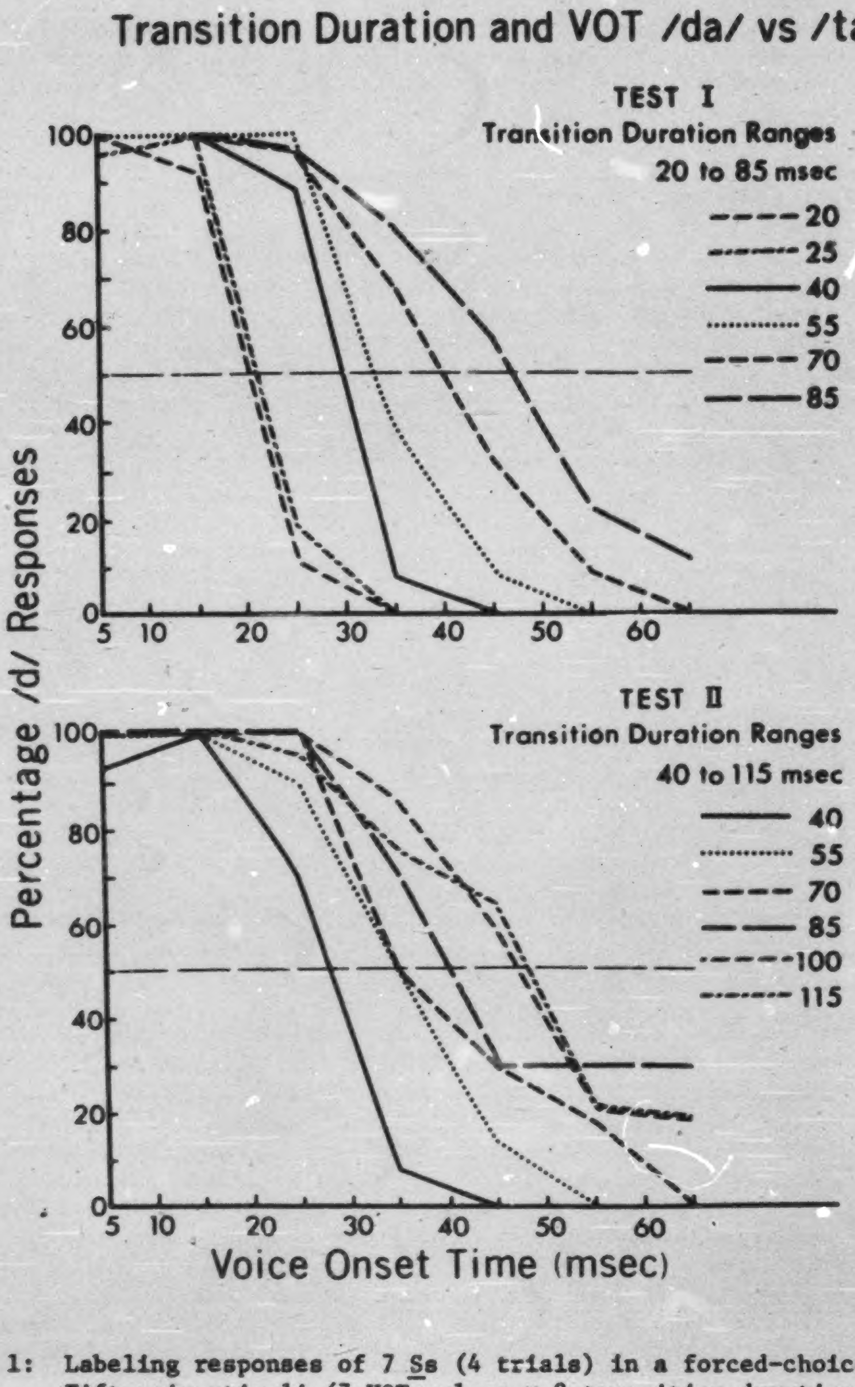

Figure 1: Labeling responses of 7 Ss (4 trials) in a forced-choice task. Fifty-six stimuli (7 VOT values × 8 transition durations) with bursts and transitions appropriate to /da/ and /ta/ were presented in two tests to reduce subject fatigue.

Now let us ask again whether these data, and the Stevens-Klatt data as well, point unequivocally to the duration of the voiced rather than the unvoiced transition as the feature that determines listeners' labeling judgments. In Figure 2 we have represented schematically the first-formant trajectories of our test stimuli. For each transition duration, at VOT = +5, the first-formant frequency rises linearly from an onset of 154 Hz to a steady-state value of 769 Hz. Since in general the first-formant intensity is zero until the periodic source in our synthesizer is turned on, for VOT values greater than +5 msec, the actual onset frequency of F_1 depends directly on VOT. Thus for a transition of 20-msec duration, the F_1 onset frequency at the VOT crossover value is about 620 Hz. In the display, the F_1 trace is a solid line to the right of the VOT crossover. To the left of that value the dashed line indicates the absence of acoustic energy at the F_1 frequency, while the higher formants, not shown, are excited by the random noise source of the synthesizer. We see that with increasing transition duration not only is there a rightward shift in VOT crossover, but also that there are changes in F_1 onset frequency and in the duration of the transition following onset of the periodic excitation. These relations are more easily seen in Figure 3. The upper panel indicates how F_1 onset frequency, or alternatively the extent of F_1 shift following voice onset, varies at the VOT boundary with changing transition duration. Given the limitations of the experiment, the two curves of course say exactly the same thing, and pending further work we cannot say which measure is more relevant perceptually, or indeed how much meaning either of them has independently of the purely temporal measures of voicing. Perhaps we might for now suppose that as VOT is increased, either the F_1 onset frequency must be lowered or the extent of its frequency shift increased in order to achieve a stimulus that is ambiguous, as between /da/ and /ta/; that is, one or the other of these changes serves to counterbalance the devoicing effect of prolonging the delay in voice onset.

The lower panel of Figure 3 suggests an answer to the question of whether the measure of VOT (here labeled "VTD" for "voiced transition duration") proposed by Stevens and Klatt (1974) provides a more stable index of stop voicing than does the usual measurement of VOT. If in fact it is true that listeners pay more attention to the transition following voice onset than to the preceding voiceless interval, then their measure ought to yield a curve of smaller slope than the standard VOT measure. This is clearly not the case here. We conclude, with Stevens and Klatt, that VOT is not alone sufficient to explain our listeners' behavior, but that VTD, their proposed measure, is even less adequate, by itself, to account for that behavior.

So far we have been talking about formant transitions as though only the first formant deserves attention in a discussion of stop voicing. To see whether this is justifiable, we ran a second experiment in which, along with VOT, the transition duration of the first formant was varied independently of the two higher formants. Transition durations of 15 and 80 msecs were used, and they were assigned to our test stimuli as shown on the left-hand side in Figure 4. The labeling data on the right-hand side of Figure 4 show that listeners' responses were determined almost entirely by the F_1 transition. With a short F_1 transition the effect of varying the higher formants is nil. With the longer F_1 transitions the higher formants have some effect, but that effect, measured as a shift in VOT crossover, is considerably smaller than the effect of a change in F_1. The effect of transition duration on stop voicing is then primarily attributable to the first formant.

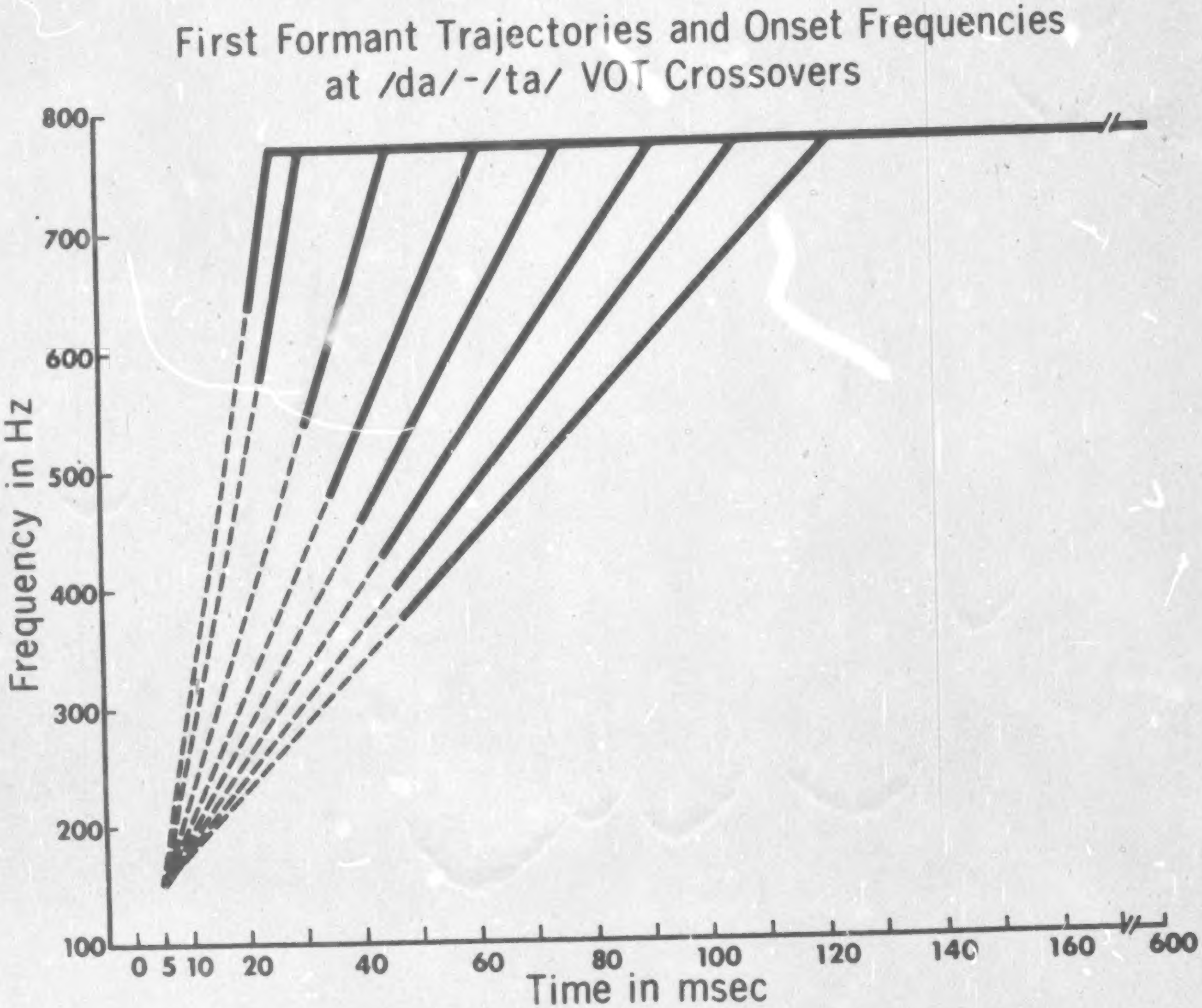

Figure 2: First-formant configurations for the 8 transition durations represented in the data of Figure 1. The values along the abscissa represent VOT. For each transition, duration values of VOT and F_1 onset frequency yielding mainly /da/ responses are represented by the dashed line; the solid line represents values yielding better than 50 percent /ta/.

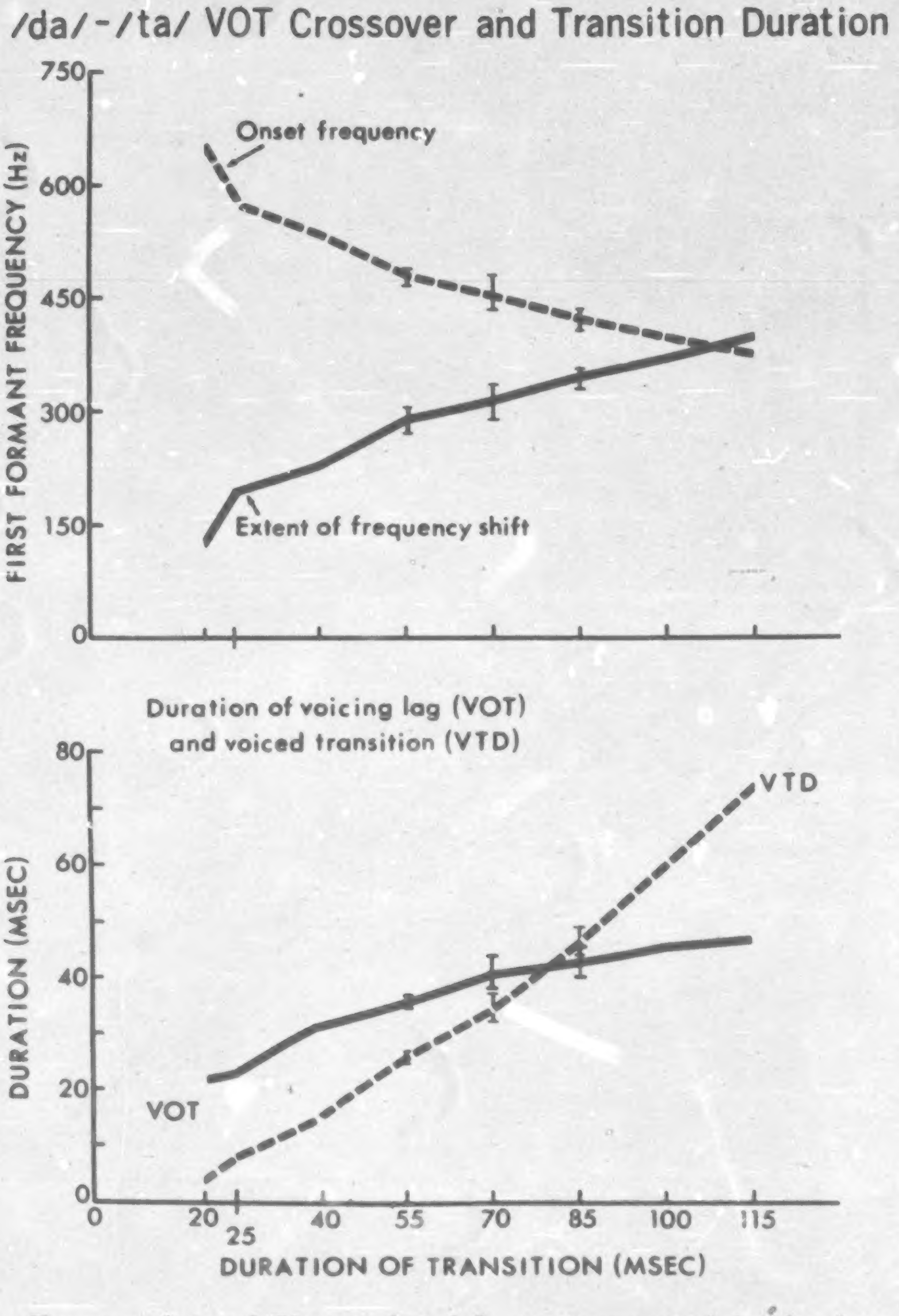

Figure 3: The same data of Figures 1 and 2 are represented in the four curves shown. For the transition durations tested twice that yielded different VOT crossover values the curves show overall mean values; the short vertical lines indicate the magnitude of the differences between Test I and Test II data.

Transition Durations of F_1 and Higher Formants: Effects on VOT Crossovers for Initial /da/ vs /ta/

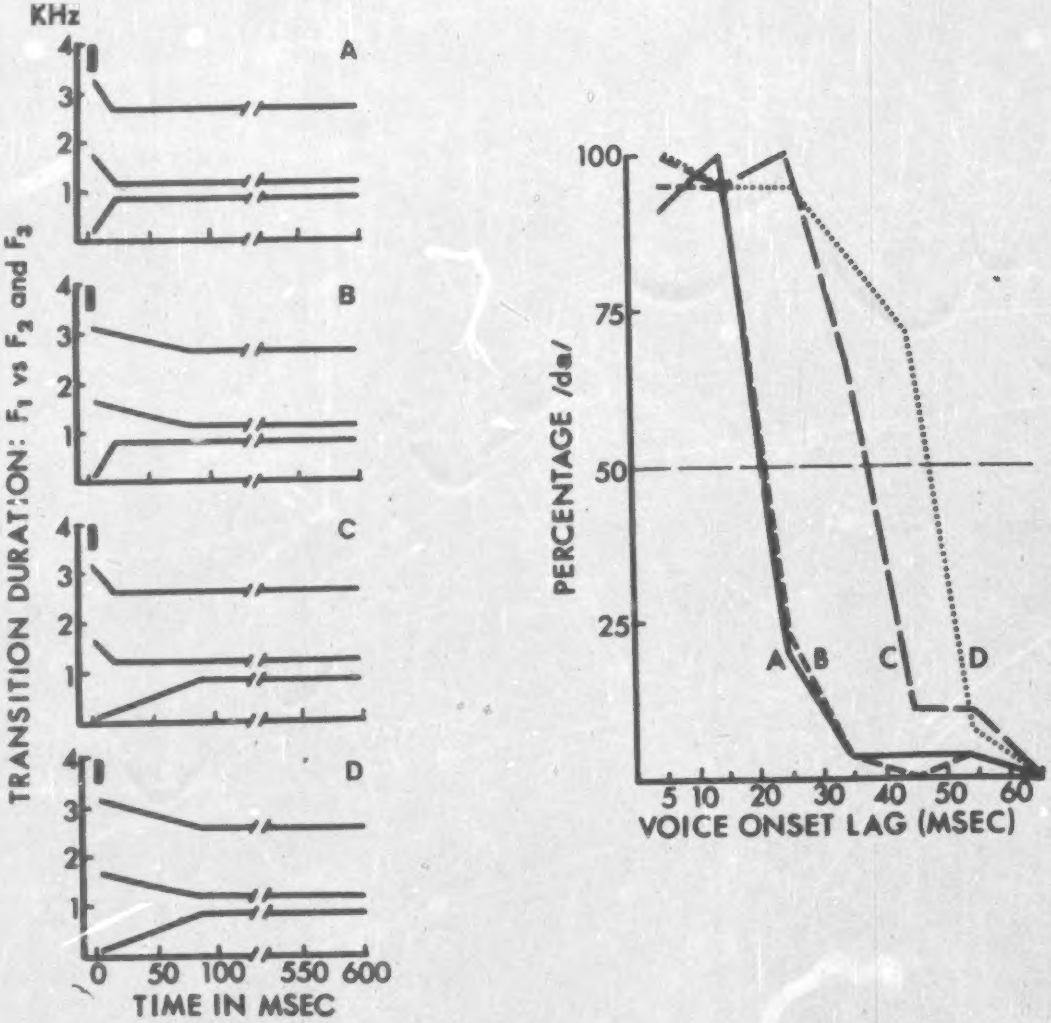

Figure 4: Patterns A and B share F_1 transitions of 15 msec, while C and D F_1 transitions are of 80-msec duration. Patterns A and C are similar in having F_2 and F_3 transitions of 15-msec duration; B and D F_2 and F_3 transitions are 80 msec long.

263
FIGURE 4

To conclude then, the VOT boundary is not fixed; it varies directly with the transition duration. However, it is limited, appearing to lie between limits of roughly 20 and 50 msec following the burst onset. The duration of the voiced transition at the boundary between /da/ and /ta/ also varies with transition duration, and our data fail to give any indication of a limiting value for this feature beyond which /ta/ could not be heard.

REFERENCES

Cooper, F. S., P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman. (1952) Some experiments on the perception of synthetic speech sounds. *J. Acoust. Soc. Amer. 24*, 597-608.

Fujimura, O. (1971) Remarks on stop consonants: Synthesis experiments and acoustic cues. In *Form and Substance: Phonetic and Linguistic Papers Presented to Eli Fischer-Jørgensen*, ed. by L. L. Hammerich, R. Jakobson, and E. Zwirner. (Copenhagen: Academisk Forlag).

Haggard, M., S. Ambler, and M. Callow. (1970) Pitch as a voicing cue. *J. Acoust. Soc. Amer. 47*, 613-617.

Liberman, A. M., P. C. Delattre, and F. S. Cooper. (1958) Some cues for the distinction between voiced and voiceless stops in initial position. *Lang. Speech 1*, 153-167.

Stevens, K. N. and D. H. Klatt. (1974) Role of formant transitions in the voiced-voiceless distinction for stops. *J. Acoust. Soc. Amer. 55*, 653-659.

Some Maskinglike Phenomena in Speech Perception*

M. F. Dorman,[+] L. J. Raphael,[+] A. M. Liberman,[++] and B. Repp[+++]

ABSTRACT

To study maskinglike effects in the perception of continuous speech, listeners were presented two-formant syllables /bɛb/ or /bɛg/ followed, at intervals from 0 to 150 msec, by /dɛ/. The subjects were instructed to identify the syllable-final consonant. An 80-msec intersyllable interval was required for recognition of the syllable-final consonant to reach asymptote. To determine the level at which this "effect" occurred, we repeated the experiment, but with the second-formant transitions of the first syllable presented alone for judgment as rising or falling chirps. Recognition of the isolated transitions (chirps) was essentially unaffected. These data suggest that the "masking" in the initial study was due to the elimination of a necessary cue--in this case, a silent interval, corresponding to the stop closure between the syllables--and not to backward masking of the auditory information. A third study found that changing voice between the syllables from male to female also eliminated almost all the "masking." This reinforces the conclusion of the first two studies, indicating in this case, that the effect is not to be attributed to interruption of information processing.

It has been suggested recently that forward and backward masking may constrain the perception of phonetic segments. Thus, Massaro (1972) has proposed that "...the redundancy of a vowel in normal speech...protects it from later speech until processing has been completed" and, in the same spirit, that "...if a consonant-vowel transition was followed by a speech sound that could not be integrated with it, perception should be disrupted, and backward recognition masking should occur."

---

*This is a revised version of a paper presented at the 89th meeting of the Acoustical Society of America, Austin, Texas, 7-11 April 1975.

[+] Also Herbert H. Lehman College of the City University of New York.

[++] Also University of Connecticut, Storrs, and Yale University, New Haven, Conn.

[+++] University of Connecticut Health Center, Farmington, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-42/43 (1975)]

In this paper we will examine several cases of speech perception that fit the paradigms for forward and backward masking. Our purpose is to see if the underlying processes are, indeed, those of masking, either peripheral (integration of target and mask) or central (disruption of processing).

For the case of forward masking, we have chosen an instance of a type long familiar to students of speech perception: to perceive the stop in a syllable-initial fricative-stop cluster, we must have a period of silence between the fricative noise and the start of the stop transitions. The syllables in our experiment are /ʃpɛ/ and /ʃkɛ/. In Figure 1 we have shown a schematic spectrogram sufficient for the perception of /ʃpɛ/. Seen in terms of the forward-masking paradigm, the noise of the fricative /ʃ/ would be the mask and /pɛ/ (or /kɛ/) the target. In our first experiment we varied the silent interval (let us call it the interstimulus interval or ISI) between the /ʃ/ mask and the /pɛ/ or /kɛ/ targets. The resulting stimulus patterns were randomized and presented to 14 listeners for judgment as /ʃpɛ/, /ʃkɛ/, or /ʃɛ/. We see in Figure 2 that "masking" did occur: at ISIs of 20 msec or less the listeners reported hearing /ʃɛ/, not /ʃpɛ/ or /ʃkɛ/.

To gain some insight into the processes underlying the failure to hear the stop, we undertook a second experiment to determine if there was, in fact, a masking of the essential acoustic cue--the second-formant transition--for the perceived distinction between /ʃpɛ/ and /ʃkɛ/. For that purpose, we followed exactly the procedures of the first experiment except that, in this case, the targets were not the syllables but only their second-formant transitions. These isolated transitions are not heard as speech; rather, they sound like "chirps," which our subjects easily learned to identify as "high" or "low." The outcome of this second experiment is shown in Figure 3. We see that correct perception of the "chirps" was not noticeably affected by the /ʃ/ mask. From that we infer that our subjects' failure to hear the stops in the first experiment was not due to masking in the ordinary sense. That is, the role of the silent interval between the /ʃ/ noise and the stop transitions is not, apparently, to avoid interference between target and mask. For a more reasonable interpretation we should note that in the production of initial fricative-stop clusters, closure must occur after the fricative, and therefore we may suppose that the silence caused by the closure is an essential manner cue for the perception of the stop. On that interpretation, the silence provides information, not freedom from interference.

Let us turn now to the paradigm for backward masking and in that connection consider the perception of the disyllables, /bɛb dɛ/ and /bɛg dɛ/. A schematic spectrogram sufficient for the perception of /bɛb dɛ/ is shown in Figure 4. As a case for backward masking, the syllable-final consonant /b/ in /bɛb/ is the target and the syllable /dɛ/ is the mask. To determine, then, whether masking does occur in this case we varied the silent interval between the mask /dɛ/ and the targets /bɛb/ or /bɛg/, randomized the resulting patterns, and presented them to 13 subjects for judgment as /bɛb dɛ/, /bɛg dɛ/, /bɛ dɛ/. The outcome is shown in Figure 5. We see that at ISIs of 50 msec or less the subjects reported hearing /bɛ dɛ/--that is, they did not hear the syllable-final stops /b/ and /g/.

To find out more about the underlying processes, we carried out for this paradigmatic case of backward masking an experiment analogous to the forward-masking experiment with the chirps. That is, we isolated the acoustic cue for

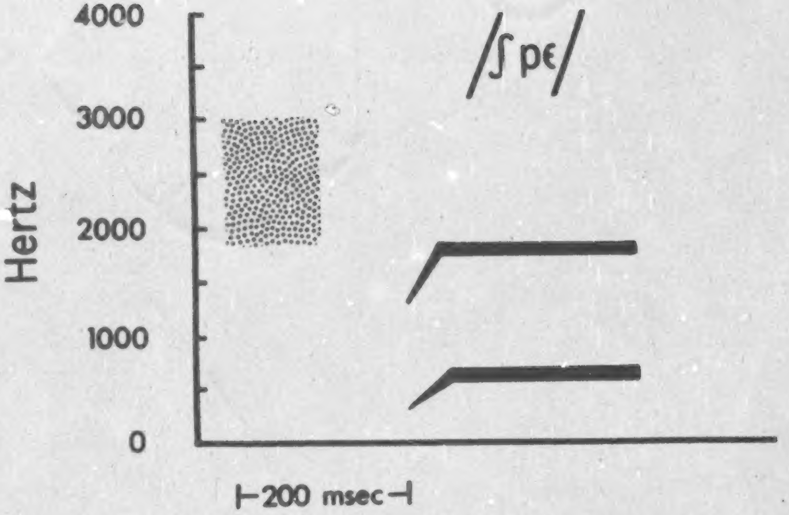

Figure 1: Stylized spectrogram of /ʃpɛ/.

FIGURE 1

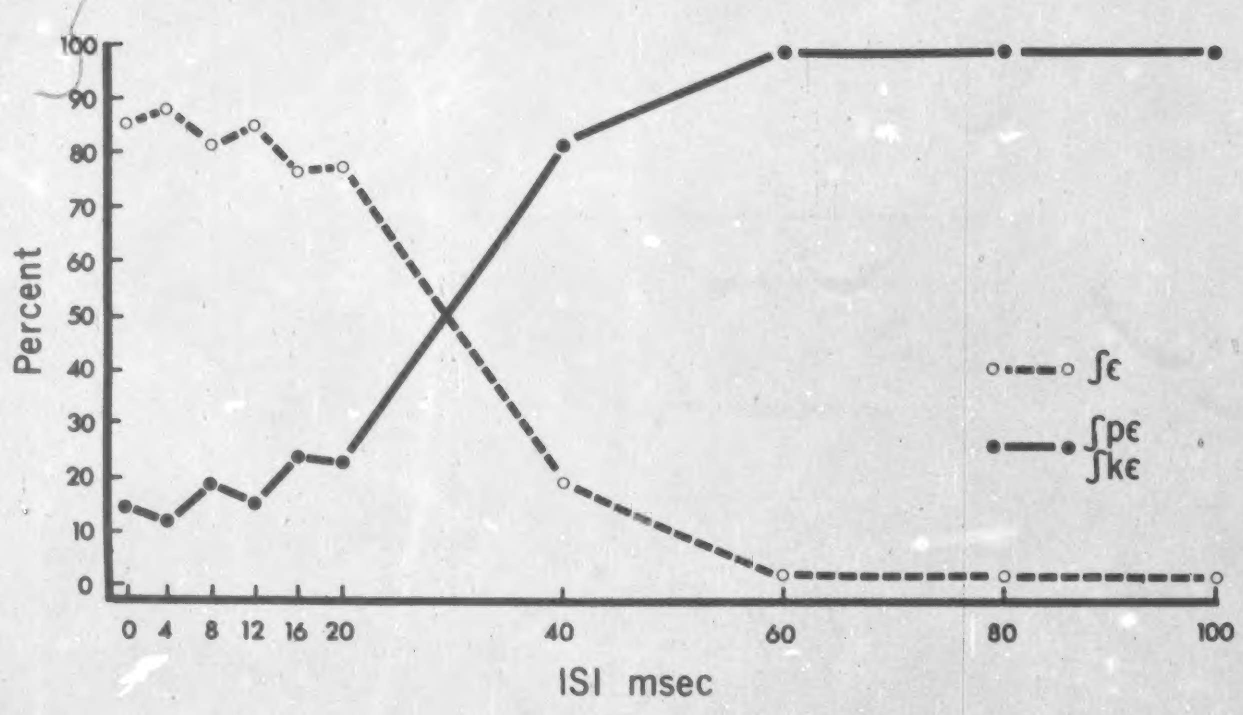

Figure 2: Percent identification of /ʃɛ/ and /ʃpɛ ʃkɛ/.

FIGURE 2

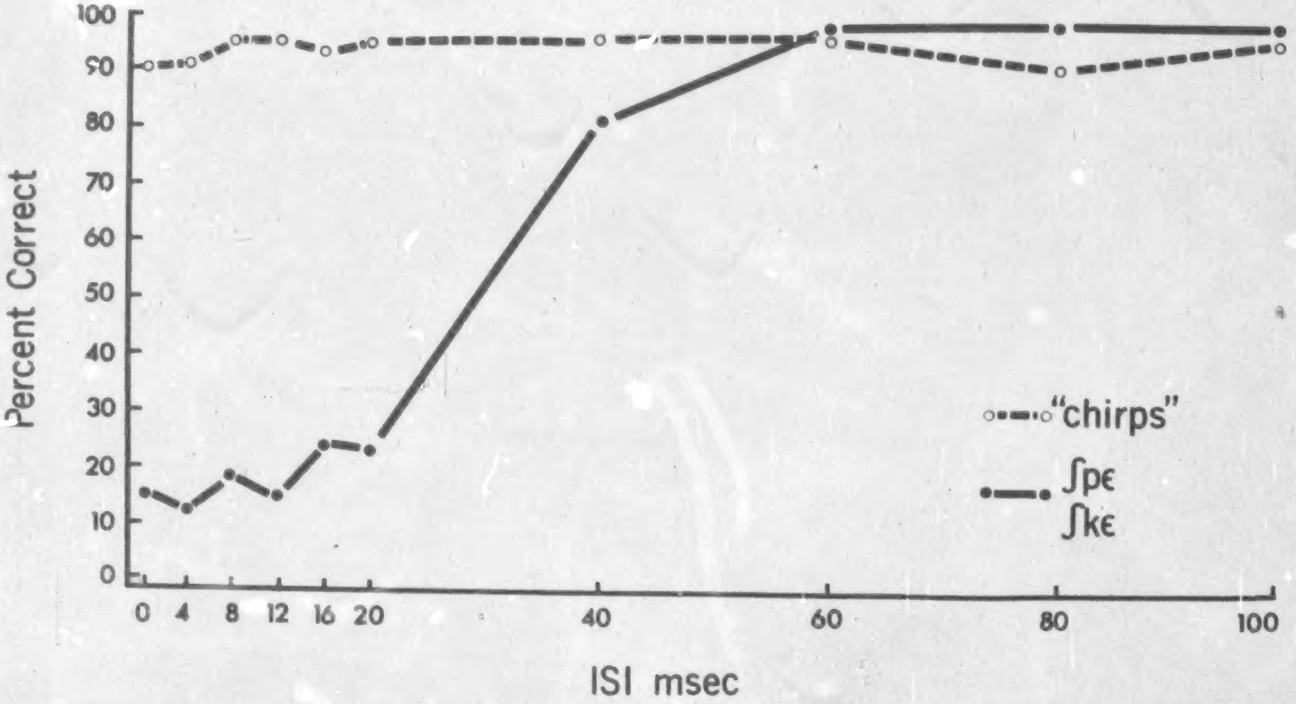

Figure 3: Percent correct identification of "chirps" and /ʃpɛ ʃkɛ/.

FIGURE 3

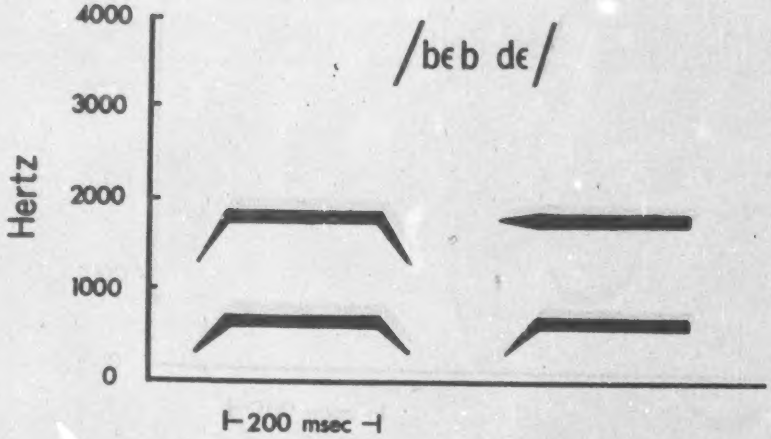

Figure 4: Stylized spectrograms of /bɛb dɛ/.

FIGURE 4

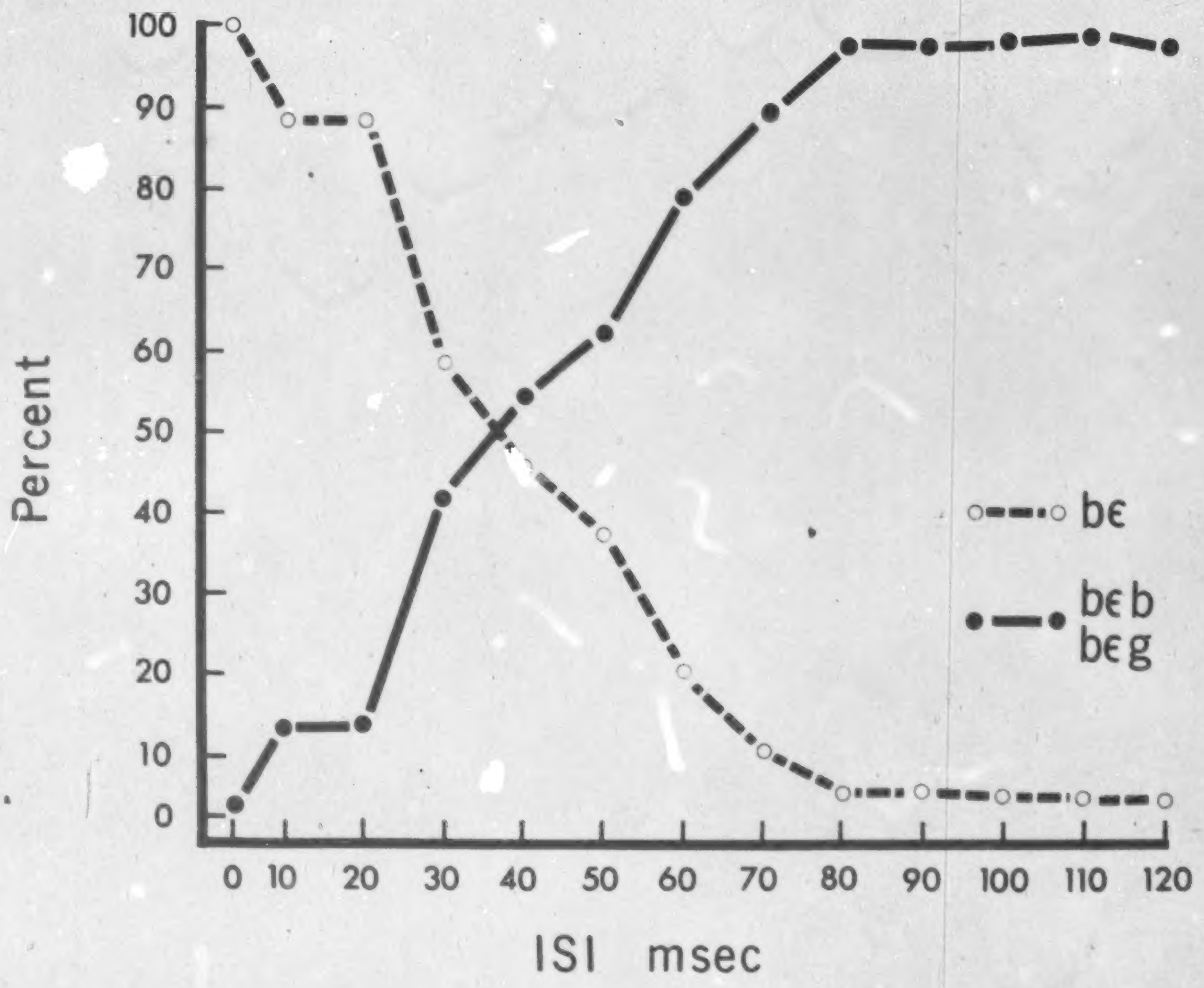

Figure 5: Percent identification of /bɛ/ and /bɛb bɛg/.

271

FIGURE 5

the perceived distinction between /bɛb/ and /bɛg/--the second-formant transitions that, by themselves, sound like chirps--and, substituting them for the target syllables /bɛb/ and /bɛg/, we added the /dɛ/ mask exactly as it had been added in the experiment where it has, at short ISIs, effectively masked the stop-consonant targets. (The subjects were taught to identify the "chirp" as high or low.) The outcome is shown in Figure 6. We see that the correct perception of the chirps was little affected by the mask. This suggests that the processes underlying the failure to hear the stops was not due to masking of the differential acoustic cue.

Once again we might suppose that, as in the paradigmatic case of forward masking, the role of the necessary silent interval is to provide information, not time for processing. As in the earlier case of forward masking, that supposition is reasonable on the basis that the disyllables /bɛb dɛ/ and /bɛg dɛ/ can be produced only if the speaker closes his vocal tract (thus producing an interval of silence) between the end of the first syllable and the beginning of the second. In the case of backward masking, however, there remains a masking interpretation to which the results with the chirps are not necessarily relevant: it is possible that the mask (the /dɛ/ syllable) interrupted the phonetic (as opposed to auditory) processing of the target. It is difficult to test that hypothesis directly, but the data of the next experiment do bear on it.

The next experiment was like the backward-masking case just described except that there were three targets--/bab/, /bad/, and /bag/--followed by a single mask /da/. The outcome is shown in Figure 7, where we see that the length of the necessary silent interval is different for the three targets; /bad/ in particular stands out, needing a much longer silent interval than the other two. One can think of no reason why the perception of syllable-final /d/ should require so much more processing time than the other stops, which is the assumption that a masking-process interpretation demands. On the other hand, it is quite obvious that the normal production of the geminate /bad da/ requires a longer vocal-tract closure than /bab da/ or /bag da/ (Delattre, 1971). Thus, this result points once again to the conclusion that silence is here a cue, a source of information. Moreover, it suggests, more compellingly perhaps than the earlier experiments, that phonetic perception is in this case constrained not primarily by what the auditory system can do but by what vocal tracts can do. The auditory system could hear the chirps even at short ISIs, but no vocal tract can produce the geminated stops with such short closures.

On the assumption that perception might here be obeying vocal-tract constraints, we ask next: Whose vocal tract? Common sense suggests that it would not be that of the listener or the speaker nor yet of any other individual, but rather some very abstract conception, which somehow takes account of what can and cannot be done by vocal tracts in general. In that connection we note that, as we have already remarked, a vocal tract cannot produce /bab da/ or /bad da/ without closing between the syllables and, as we have found, a listener cannot perceive both stops unless there is a corresponding interval of silence. But that applies only to a single vocal tract. Given two speakers, one can produce the first syllable and the other the second syllable with no silent interval between. We thought it of some interest to determine experimentally if the constraints that applied in the perception of utterances produced by a single voice would apply equally when perceptibly different voices produce the target syllable and the mask syllable.

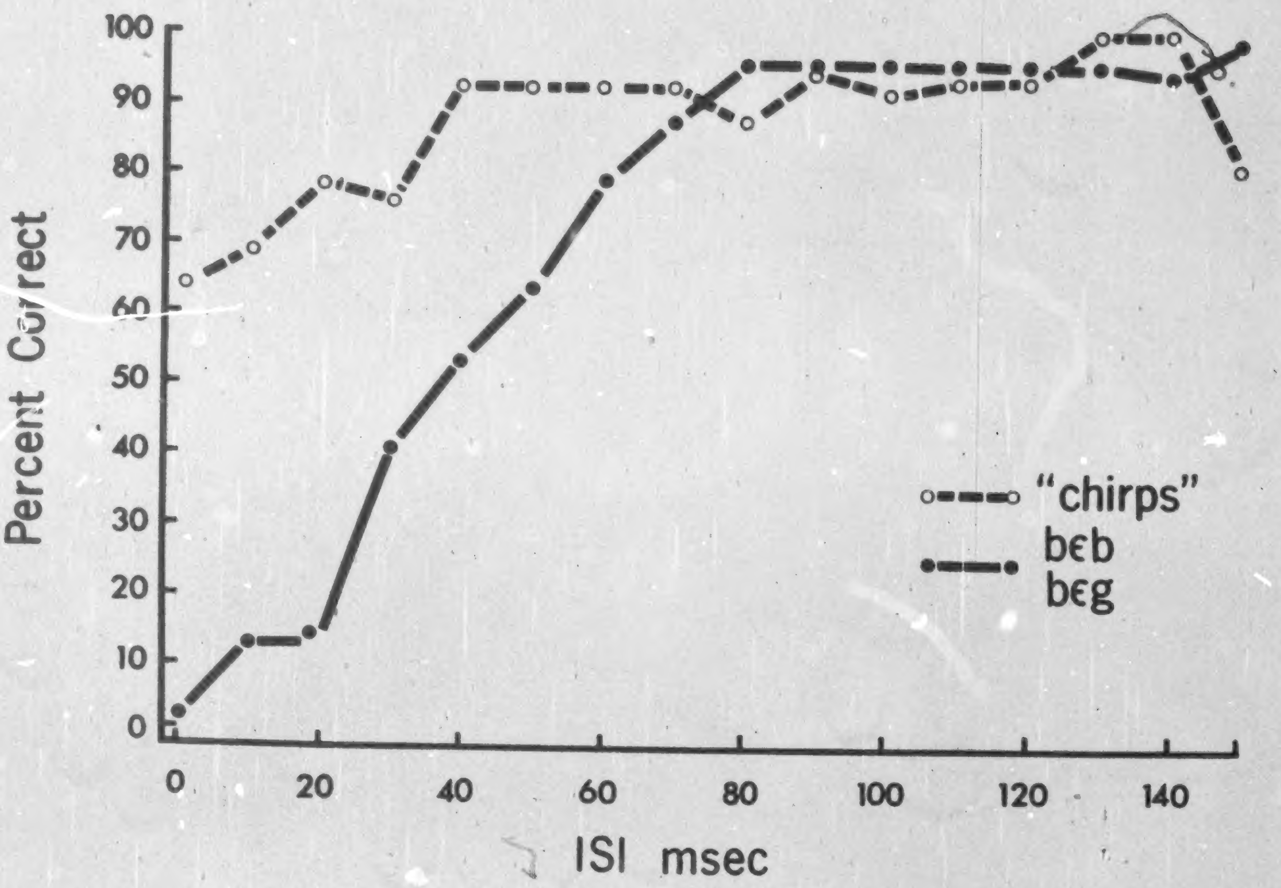

Figure 6: Percent correct identification of "chirps" and /bεb bɛg/.

FIGURE 6

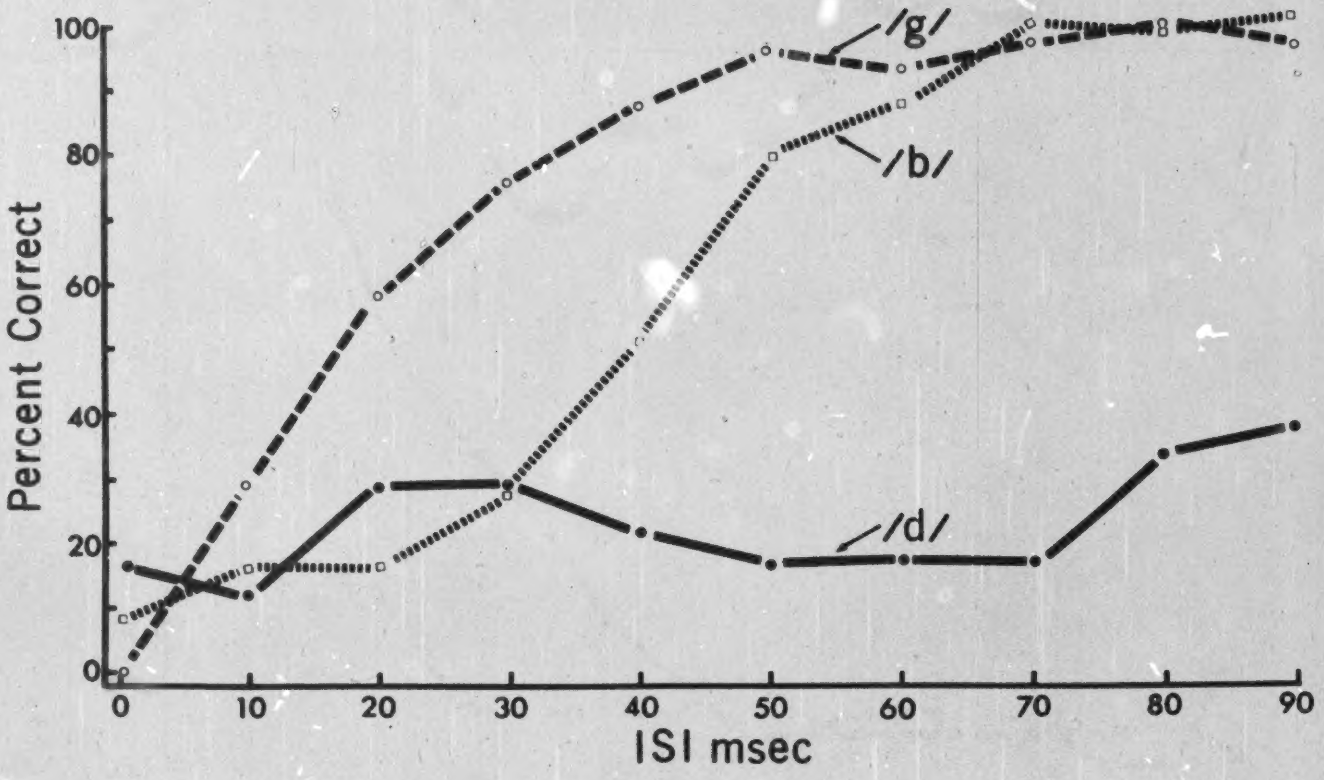

Figure 7: Percent correct identification of syllable-final /b d g/ when followed by /d /.

FIGURE 7

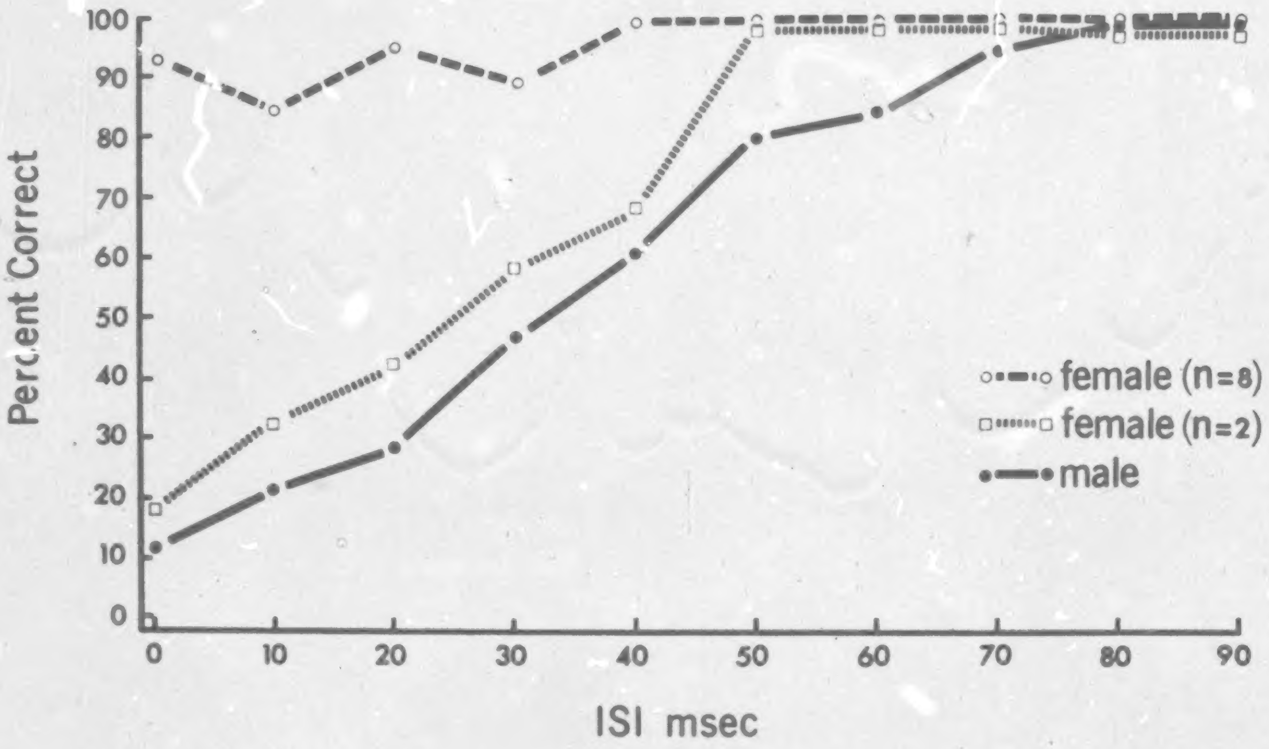

Figure 8: Percent correct identification of syllable-final stops when followed by a syllable in the same voice or in a different voice.

FIGURE 8

In the last experiment, then, we duplicated the procedures of the earlier experiment in which we had varied the silent interval between target syllables /bab/ and /bag/ and a mask /da/, but in one case both syllables were spoken by a male while in the other the target was spoken by a male and the mask by a female. Two test sequences were produced, one for the same-voice condition, the other for the different-voice condition. Within each sequence the patterns were randomized and presented for judgment of the syllable-final consonant /b/ and /g/, as in the earlier experiments. Ten subjects were told before each condition that the syllables either were or were not produced by the same voice. The outcome is shown in Figure 8. The solid line (labeled "male") is for the same-voice condition. We see much the same result that had been found earlier: at relatively short ISIs the syllable-final consonants are not heard. The results for the different-voice condition are shown by the dashed lines (labeled "female"). We see that eight of the listeners in the different-voice condition correctly perceived the syllable-final consonants, even at the very shortest ISIs. Two of the listeners performed in the different-voice condition exactly as they had in the same-voice condition. (It is worthy of note that one of these two listeners spontaneously commented at the end of the experiment that she had not thought that the voices were different; she had assumed, rather, that it was the same person speaking at two different pitches.)

Though many controls need now to be carried out, we shall tentatively conclude on the basis of the last experiment that if phonetic perception is, in any case, constrained by what vocal tracts can and cannot do, the constraint is a very abstract one indeed. From all the experiments here reported, we shall conclude, somewhat less tentatively, that the role of silence before (or after) a stop is not to avoid interference (as between target and mask) but to provide important information. Putting that less tentative conclusion together with the more tentative one, we might say that the information is important because it tells the listener what a vocal tract is doing.

REFERENCES

Delattre, P. (1971) Consonant gemination in four languages: An acoustic, perceptual and radiographic study. Part I: International Review of Applied Linguistics 9, 31-52; Part II: International Review of Applied Linguistics 9, 97-113.

Massaro, D. (1972) Preperceptual images, processing time, and perceptual units in auditory perception. Psychol. Rev. 79, 124-145.

The Perception of Vowel Duration in VC and CVC Syllables*

Lawrence J. Raphael,+ Michael F. Dorman,+ and A. M. Liberman++

ABSTRACT

It is well-established that consonant-vowel (CV) transitions simultaneously convey information about the consonant and the vowel. One wonders then whether listeners can perceive separately the durations of consonant and vowel. This question is of some consequence, since it is known that "vowel" duration is a cue for the perceived distinction between unreleased voiced and voiceless stops in final position in the syllable. Our aim in this study was to find out whether transitions convey information about vocalic duration. In order to answer this question, three pairs of synthetic stimuli were generated. One member of each pair was a vowel-consonant (VC) syllable, the other was a consonant-vowel-consonant (CVC) syllable. All stimuli ended in transitions appropriate to unreleased /d/, and each contained formants, appropriate to the vowel /ɛ/, which were varied in 10-msec steps over a durational range of from 30 to 150 msec. The initial consonants of the CVC stimuli in each of the three experiments were, respectively, /d/, /r/, and /s/. Listeners were asked to identify the final consonant as voiced or voiceless (i.e., as /d/ or /t/). For steady-state vowels of equal duration, the presence of an initial consonant caused a shift in the perceptual boundary between the /t/ and /d/ categories, relative to the boundary found for the VC stimuli. The shift was found to be greater both for a greater duration of transitional input (e.g., /r/ vs. /d/) and for vocalic transitions, as opposed to noise (e.g., /r/ vs. /s/). This outcome suggests that a portion of the initial consonant transition duration is used by the listener in estimating vowel duration, at least for the purpose of cuing the voiced/voiceless distinction in final position, and thus that there is no discrete perception of the durations of the vowel or of the consonant in the syllable.

---

*This is a revised version of a paper presented at the 89th meeting of the Acoustical Society of America, Austin, Tex., 7-11 April 1975.

+Also Herbert H. Lehman College of the City University of New York.

++Also University of Connecticut, Storrs, and Yale University, New Haven, Conn.

Acknowledgment: We acknowledge the assistance of Tony Levas and Suzi Pollock in the data analysis.

[HASKINS LABORATORIES: Status Report on Speech Research SR-42/43 (1975)]

INTRODUCTION

Many studies have demonstrated the importance of formant transitions of vowels as cues to the perception of consonants (cf. Liberman, 1957; Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). Moreover, it has been shown that formant transitions simultaneously carry a variety of perceptual cues to consonant identity. For example, consonantal manner information (rate of change of F_1), place-of-articulation information (locus of F_2 and F_3), and voicing information (degree of F_1 attenuation) are transmitted in parallel on formant transitions.

Although formant transitions are commonly referred to as "consonant transitions" because of the nature of the cues they carry, they are, nevertheless, changes in frequency of the formant structures that characterize vowels. Thus, formant transitions carry in parallel both consonant and vocalic information. The purpose of the present study was to investigate to what extent, if any, transitions of vowel formants contribute information about vowel duration, and about the duration of the syllable whose nucleus is the vowel.

EXPERIMENT I

Several studies have shown that vowel duration may be used to cue the voicing characteristic of consonants in syllable-final position (Denes, 1955; Raphael, 1972; Raphael, Dorman, Tobin, and Freeman, in press). Varying the steady-state vowel duration in a synthetic VC syllable speech, as /ɛd/, will cause listeners to perceive the final consonant as /t/ or /d/: relatively short steady-state vowel durations cue /t/ judgments; relatively long vowel durations cue /d/ judgments. If formant transition duration is perceptually accessible for incorporation with the steady-state portion of the vowel, then the effect of adding initial consonant transitions to the VC syllable should be to shift the /t/-/d/ phoneme boundary toward the shorter end of the vowel range; that is, more final consonants should be identified as voiced.

Method

The experimental stimuli were produced on the Haskins Laboratories parallel resonance synthesizer. One set of stimuli consisted of a three-formant vowel, /ɛ/, followed by 50-msec formant transitions appropriate for the voiced stop consonant /d/. The stop was synthesized without cues for release. The steady-state duration of the vowel was varied in 10-msec steps from 30 to 150 msec. Another stimulus series used the VC stimuli as a base, but attached 60-msec formant transitions appropriate for [d] to the beginning of each VC signal. Six tokens of each of the VC and CVC stimuli were generated. These stimuli were then randomized, recorded, and reproduced for 12 beginning phonetic students at Herbert H. Lehman College. The stimuli were delivered over a loudspeaker in a sound-treated room in the audiology laboratory of the college. The listeners were instructed to identify the final consonant of each stimulus as /t/ or /d/.

Results and Discussion

The percentage of /t/ responses for both the VC and CVC stimuli are shown in Figure 1. Phoneme boundaries were determined by fitting straight-line functions to the identification functions for both VC and CVC stimuli. The difference

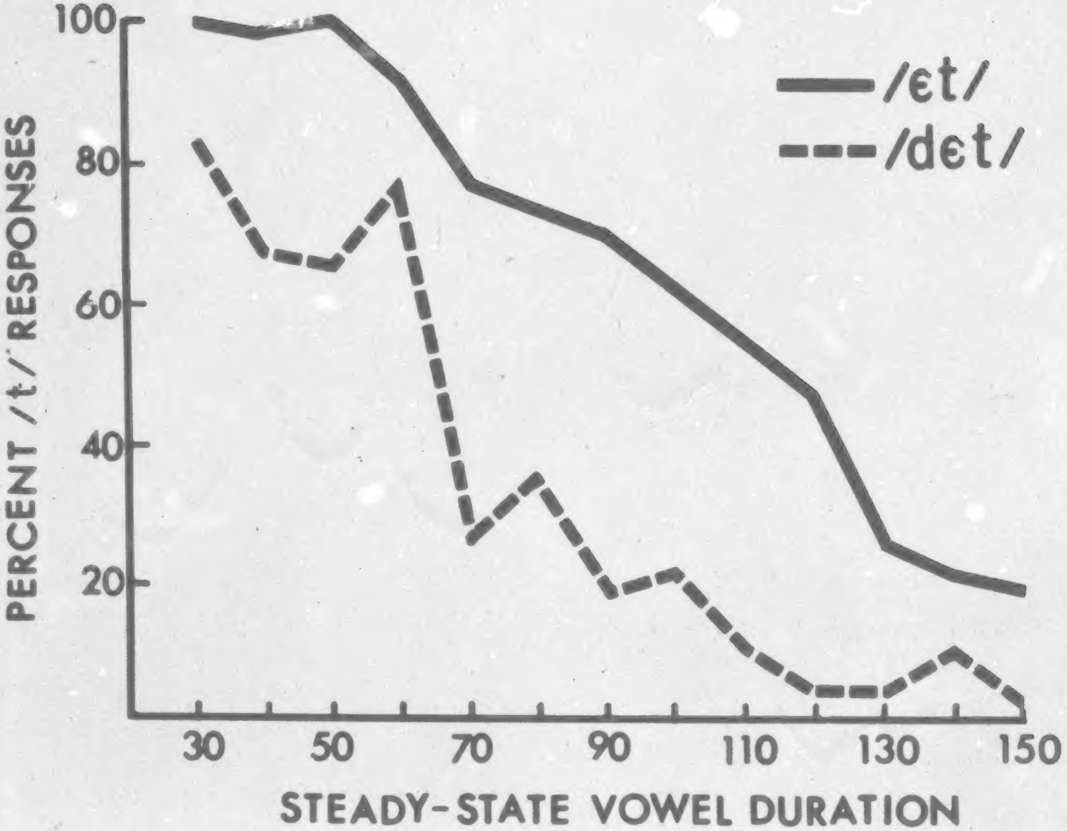

Figure 1: Percent /t/ responses for the /ɛt ɛd/ and /dɛt dɛd/ series.

FIGURE 1

in phoneme boundaries for the two series was 43 msec. This outcome indicates that at least some of the duration of the formant transitions associated with the initial consonant was processed by the listeners as part of the vowel/syllable duration.

EXPERIMENT II

Experiment I revealed that some portion of the initial formant transition duration may be used to determine vowel duration. The purpose of Experiment II was to replicate and extend the outcome of the first experiment by assessing the effect of longer (100-msec) intital formant transitions on the perception of voicing in syllable-final consonants.

Method

The VC stimuli of Experiment I were also used in Experiment II. In addition, a series of stimuli were created by attaching 100-msec duration formant transitions appropriate for /r/ to the beginning of each of the VC stimuli. The /r/ formant transitions were created by linear interpolation from the starting resonances to the vocalic resonances. Thus, the /r/ transitions did not contain initial steady-state resonances. Six tokens of each stimulus were generated. These stimuli were then randomized and recorded on audio tape. The listeners of Experiment I also participated in Experiment II. As in Experiment I, the listeners were instructed to label each stimulus as ending in /t/ or /d/.

Results and Discussion

The percentage of /t/ responses for the CVC and VC stimuli are shown in Figure 2. The phoneme boundaries differed by 65 msec. As in the first experiment, a shorter duration of steady-state vowel was associated with the /t/-/d/ boundary in the CVC syllables, again suggesting that at least some of the duration of the transitions for the initial /r/ was incorporated into the duration of the vowel/syllable.

EXPERIMENT III

Since the 100-msec /r/ transition caused a greater shift in the phoneme boundary than did the 60-msec /d/ transitions, we may suppose that listeners are able to incorporate duration information from transitions in proportion to the magnitude of the formant transition duration. The purpose of Experiment III was to determine whether vocalic transitions, that is, resonances that predict the formant locus of the following vowel, are a necessary condition for the vowel/syllable lengthening effects of Experiments I and II. To this end, Experiment III compared the identification of syllable-final /t/-/d/ in VC stimuli, and in CVC stimuli in which the syllable-initial consonant was /s/.

Method

The VC stimuli used in Experiments I and II were also used in Experiment III. Vowel durations ranged from 50 to 150 msec in 10-msec steps. In addition, a series of CVC stimuli were created by attaching, with the aid of the Haskins' PCM system, a 100-msec natural speech /s/ to the beginning of each of the VC stimuli. As in the previous experiments, six tokens of each CVC and VC stimulus

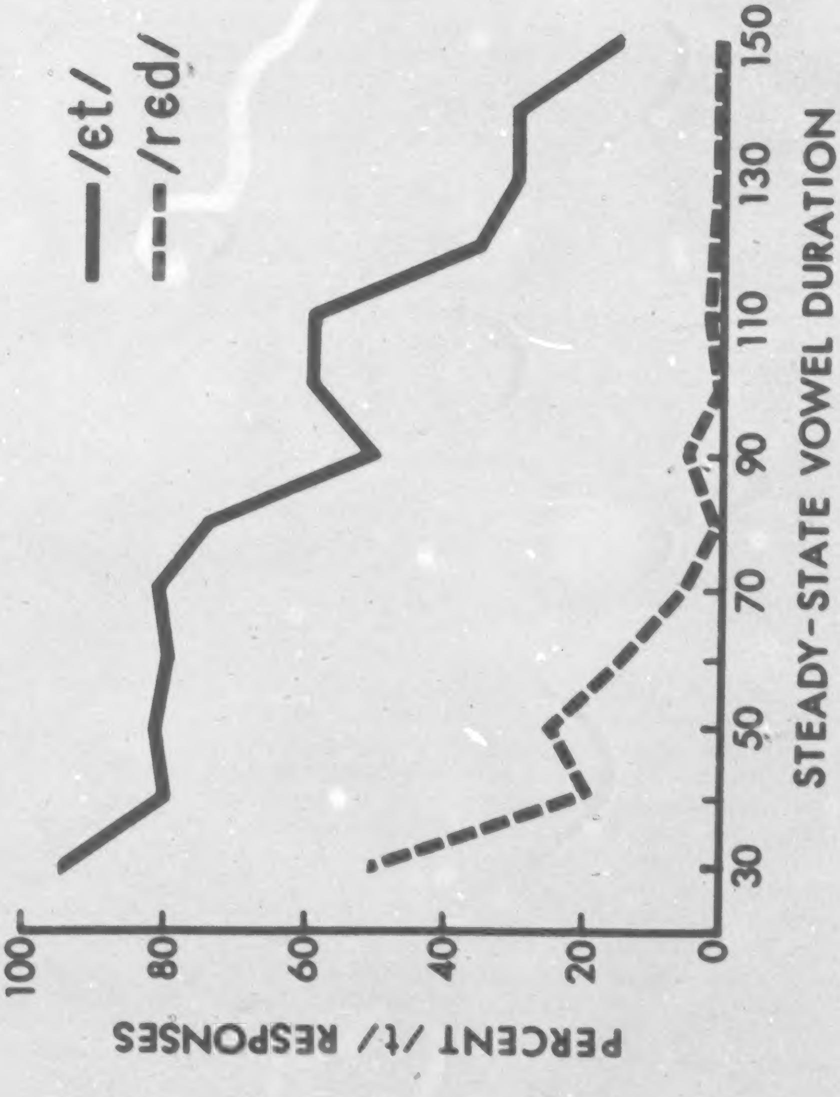

Figure 2: Percent /t/ responses for the /εt εd/ and /rεt rεd/ series.

FIGURE 2

was generated, randomized, and recorded on audio tape. The listeners were 11 beginning phonetic students at Herbert H. Lehman College. The listening conditions and task were identical to those in the first two experiments.

Results and Discussion

Figure 3 displays the percentage of /t/ responses for the VC and CVC stimuli. The phoneme boundary differed between the two conditions by 6 msec. Although the overall boundary shift was small, 10 of the 11 listeners evidenced the shift. The magnitude of the boundary shift suggests that the larger boundary shifts found in Experiment I and II were primarily due to the incorporation of the formant transition duration into the estimate of vowel duration, and were not due to increased syllable duration. The small boundary shift also suggests that well-defined formant structure may be a prerequisite for the relatively large vowel lengthening effects of Experiments I and II.

CONCLUSION

The results of Experiments I-III indicate that consonant and vowel durations in CV(C) sequences are not processed separately, but rather that they are, to a certain extent, combined to give a perceptual estimate of vowel and syllable durations. It seems clear that durational information in voiced formant transitions (e.g., /d/ and /r/) is perceptually accessible to listeners for incorporation into an estimate of vowel duration. Further, our results indicate that the longer the initial voiced formant transitions, the greater the lengthening of perceived vowel duration. However, only part of the duration of the formant transitions is used in the estimate of vowel duration. We infer this from the fact that the shift in phoneme boundaries in Experiments I and II was less than the total duration of the formant transitions. It may be that some minimal duration of formant transition information is necessary for consonant identification and that the remainder may be processed as part of the vowel. Such a conclusion, however, awaits the results of further experimentation.

The effect of the initial fricative in the CVCs of the third experiment, i.e., a small but consistent shift in the phoneme boundary, is interesting for several reasons. First, it provides an indication that syllable duration, and not just vowel duration, is operative as a cue in the final consonant voicing distinction. The data suggest, however, that the syllable duration cue is perceptually much less salient than the vowel duration cue: for an equal duration (100-msec) signal, the formant-poor /s/ provided a vary small shift in the phoneme boundary compared to the formant-rich /r/. Second, we may speculate that consonant cues conveyed by formants contribute relatively heavily to the perception of the overall duration of the vowel or syllable, whereas those contained in noise provide relatively less perceptual input to the estimate of vowel/syllable duration. Once again, conformation of such speculation awaits further research. We are currently assessing the effect of syllable-initial [h], i.e., noise-excited formants contiguous with the voiced formants of the following vowel; [t^h], i.e., transient noise with no formant structure, followed by [h]-like noise with clear formant structure that is contiguous with the following vowel; and [m], i.e., well-defined formant structure, but with a discontinuity between consonantal and vocalic resonances, on the identification of voicing in syllable-final stop consonants.

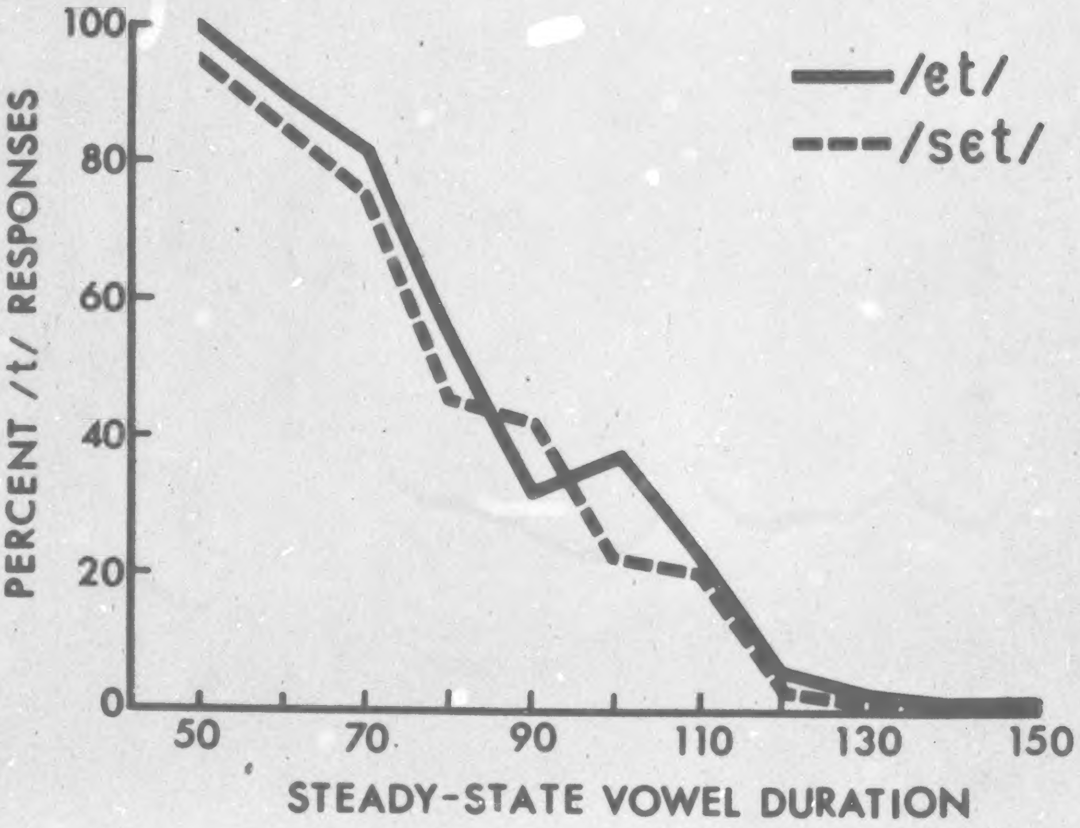

Figure 3: Percent /t/ responses for the /ɛt ɛd/ and /sɛt sɛd/ series.

FIGURE 3

REFERENCES

Denes, P. (1955) Effect of duration on the perception of voicing. J. Acoust. Soc. Amer. 27, 761-764.

Liberman, A. M. (1957) Some results of research on speech perception. J. Acoust. Soc. Amer. 29, 117-123.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-467.

Raphael, L. J. (1972) Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in English. J. Acoust. Soc. Amer. 51, 1296-1303.

Raphael, L. J., M. F. Dorman, F. Freeman, and C. Tobin. (in press) Vowel and nasal durations in vowel-nasal-consonant sequences in English: Spectrographic and perceptual studies. J. Speech Hearing Res.

On Accounting for the Poor Recognition of Isolated Vowels*

Donald Shankweiler,[+] Winifred Strange,[++] and Robert Verbrugge[+++]

ABSTRACT

Earlier studies have shown that vowels spoken in isolation tend to be poorly perceived, even when they are produced by phonetically trained talkers. Listeners, however, generally make remarkably few errors in identification of vowels in consonant-vowel-consonant (CVC) environment, even when each syllable is uttered by a different talker. Sets of nine American English vowels were spoken by a panel of talkers: in isolation and in a /p-p/ environment. Measurements of the first three formant frequencies were obtained from spectrograms. Listening tests were made up by randomizing talkers and tokens and these were presented to phonetically naive listeners. Percent recognition of the intended vowel (averaged over vowels) was 83 percent for the /p-p/ condition and 58 percent for the isolated condition. When the asymptotic formant frequencies of each talker's isolated and medial vowels are compared, the values are found to be highly similar. A nontrivial explanation must be sought for the perceptual difficulty of isolated steady-state vowels. The data point to the conclusion that no single temporal cross section of a syllable conveys as much vowel information to a perceiver as is given in the dynamic contour of the formants.

Central to current conceptions of the vowel is the idea of the target. In articulatory terms the target is a configuration of the vocal tract toward which the articulators aim. In practice, ideal vowel targets are defined in the acoustic record by formant frequency values obtained from quasi steady-state vowels produced in isolation. It is well-known, of course, that in words and sentences steady states are rarely attained and that formant frequencies usually

---

*Paper presented at the 89th meeting of the Acoustical Society of America, Austin, Texas, 10 April 1975.

[+] Also University of Connecticut, Storrs.

[++] University of Minnesota, Minneapolis.

[+++] University of Michigan, Ann Arbor.

Acknowledgment: The research reported here is a cooperative endeavor shared by the Center for Research in Human Learning of the University of Minnesota, and Haskins Laboratories. It is supported in part by grants to the Center and to Haskins Laboratories by the National Institute of Child Health and Human Development.

vary as a function of time throughout the vowel. Consonantal environment produces systematic shifts in formant frequencies causing them to deviate from target values in direction and amount that is largely predictable from articulatory considerations (Lindblom, 1963; Stevens and House, 1963). It is not known how perceivers take account of context-conditioned variation in perception of vowels. Generally it is assumed that the listener extracts the target formant values whether or not these are acoustically realized in the signal. This view encounters a difficulty, however. It has been noted several times in the literature that isolated vowels tend to be poorly perceived. This raises the question of whether vowels can be adequately described by a compact table containing only a single value for each of the first two or three formants.

In a previous study (Strange, Verbrugge, and Shankweiler, 1974), we investigated the contribution of consonantal environment to perception of medial vowels. Nine American English vowels were produced in a number of consonantal environments and in isolation by a panel of 15 untrained talkers, which included five children--ranging in age from four to ten--five adult females, and five adult males. The utterances were recorded on magnetic tape and assembled into a set of listening tests by randomly mixing the voices from token to token. We presented the tests to a group of listeners for whom these were novel voices. Misidentification of isolated vowels occurred with significantly greater frequency than of medial vowels in a number of consonantal environments. Here we present the results of perception tests for vowels in /p-p/ environment and in isolation. An average of 17 percent of the vowel nuclei were misidentified as the talker's intended vowel when they occurred in the /p-p/ frame, and 42 percent were misidentified in isolation. Figure 1 shows a vowel-by-vowel breakdown of the errors.

It is apparent that the presence of a consonantal environment produced a consistent facilitation in identification of all nine vowels. Our data are in agreement with earlier findings of Fairbanks and Grubb (1961), Fujimura and Ochiai (1963), and Lehiste and Meltzer (1973). We can conclude that isolated vowels are significantly more often misidentified than medial vowels spoken under comparable conditions. Could it be that the acoustic complexities introduced by syllabic structure better serve the requirements of the perceptual apparatus than do quasi steady-state targets?

Before accepting this conclusion, we must first ascertain whether the talkers produced isolated vowels with formant frequencies uncharacteristic of the values reported by earlier investigators. Since isolated vowels do not typically occur in natural speech, we cannot overlook the possibility that our talkers, who had no training in phonetics, produced them in peculiar ways that rendered them relatively unintelligible to the listeners.

The purpose of this study was to investigate that possibility. We undertook spectrographic analysis of the tokens of isolated vowels and medial vowels used in our listening tests. Spectrograms were made on a voiceprint spectrum analyzer of the tokens used in the perceptual tests. Sections were made at the point of closest approach to steady state. Center frequencies of F_1, F_2, and F_3 were measured. Tokens uttered by children and women were first rerecorded at half-speed to facilitate the task of locating the formants. We turn first to the results for isolated vowels.

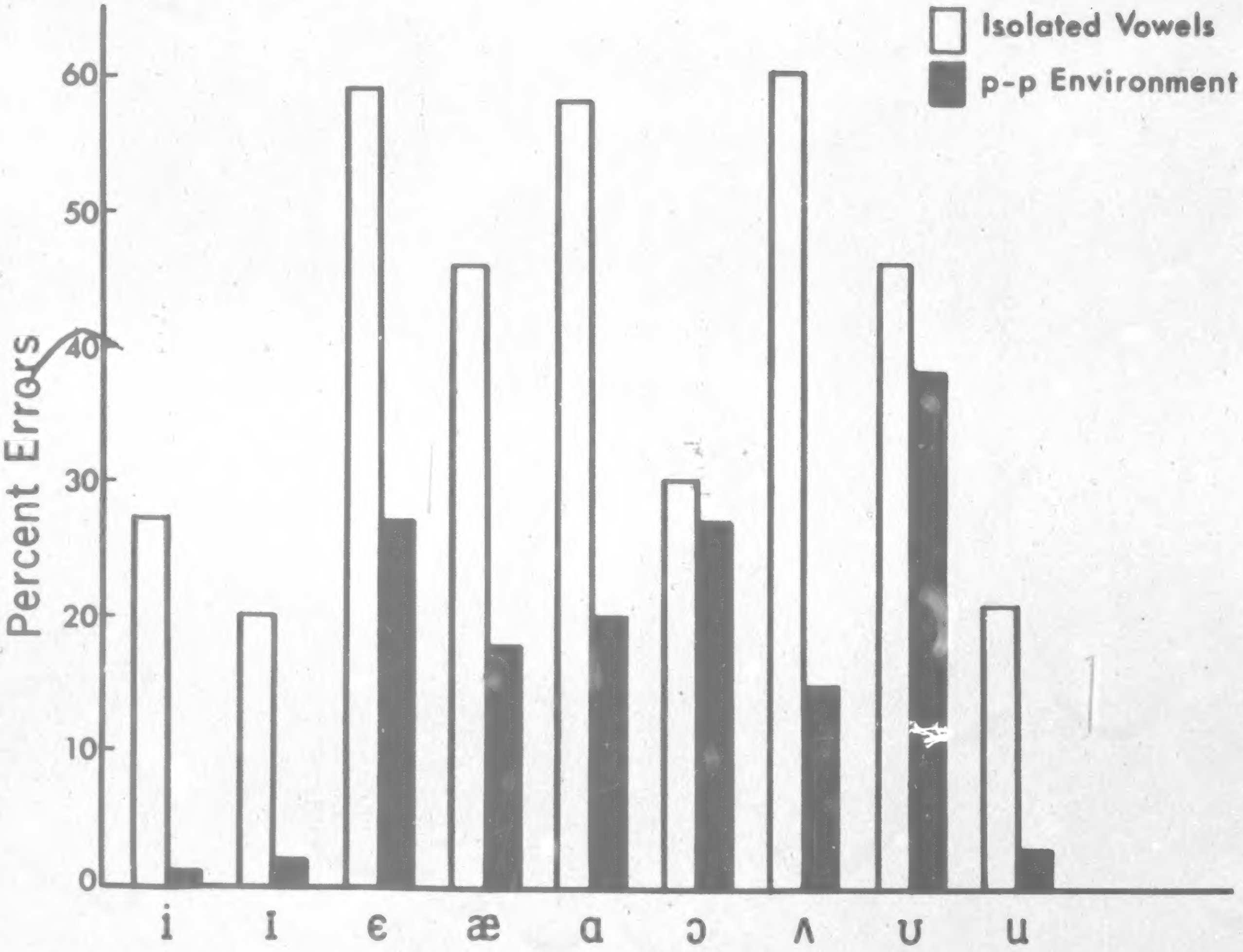

Figure 1: Mean percent errors in identification of each of nine vowels in isolation and in /p-p/ environment for 15 talkers, randomly mixed (from Strange et al., 1974).

FIGURE 1

In Figure 2, we see F_1/F_2 plots of each of five tokens of the nine vowels for the five children. The enclosed areas include all the tokens of a given type. Asterisks give the average values based on child talkers from the Peterson and Barney (1952) survey. The Peterson-Barney values make an appropriate standard of comparison inasmuch as Stevens and House (1963) have demonstrated that asymptotic formant frequencies of vowels in /h-d/ environment, as employed by Peterson and Barney, closely approximate those obtained for vowels in isolation. The figure shows that, with the exception of /æ/ and /ɔ/, our measurements cluster around the Peterson-Barney average values.

Figures 3 and 4 give the data for adult females and males, respectively. For all three groups of talkers, when the tokens are segregated by category of speaker, the vowels separate out well, except for the women's back vowels. The values for /ɔ/ tend to be displaced in the direction of /ɑ/, reflecting the predominant dialect of the upper Midwest, which minimizes the /ɑ/-/ɔ/ distinction. But, in general, it is apparent that the target values attained by our talkers in production of isolated vowels agree rather well with the values reported by Peterson and Barney (1952) for vowels in an /h-d/ environment. Thus, we can conclude that our talkers, for the most part, adopted conventional targets in their productions of isolated vowels.

Figure 5 depicts the vowel spaces for children, women, and men based on average values of F_1 and F_2 for each vowel. These vowel polygons show characteristic differences for the three categories of talkers.

The next step was to compare directly the talkers' vowel spaces for isolated vowels and for medial vowels in /p-p/ environment. Figure 6 shows that the spaces are largely congruent. F_2 of the medial vowels showed a slight migration toward the center of the space. This is in accord with results reported by Stevens and House (1963) for vowels produced between labial consonants. The effect of the shift is to reduce the acoustic contrast among the medial vowels relative to the isolated vowels.

Figure 7 displays all the tokens actually used on the listening tests arrayed in F_1/F_2 space. Isolated vowels are displayed in black; medial vowels in gray. Here it is abundantly clear that the sets of vowel tokens produced under the two conditions of this experiment occupy approximately the same space. The isolated vowels are, as expected, a little better separated acoustically than the vowels in /p-p/ environment.

It should be mentioned that for both sets of vowels a proportion of the tokens deviates markedly from the average values. Before we accept the conclusion that the inferior intelligibility of isolated vowels cannot be attributed to aberrant formant values, we should ask whether error rates on individual tokens can be predicted from their acoustic distance from an average value. If this is the case, then it becomes important to establish whether the variability of targets attained for a given vowel is greater for isolated vowels than for vowels in /p-p/ environment. A full answer to these questions awaits further study. We know, however, that the relative intelligibility of a token cannot be estimated very precisely from its position in the space defined by the two formants, a fact also noted by Peterson and Barney (1952).

Likewise, measurements of vowel duration indicate that differences in durations of medial vowels and isolated vowels fail to account for the consistent

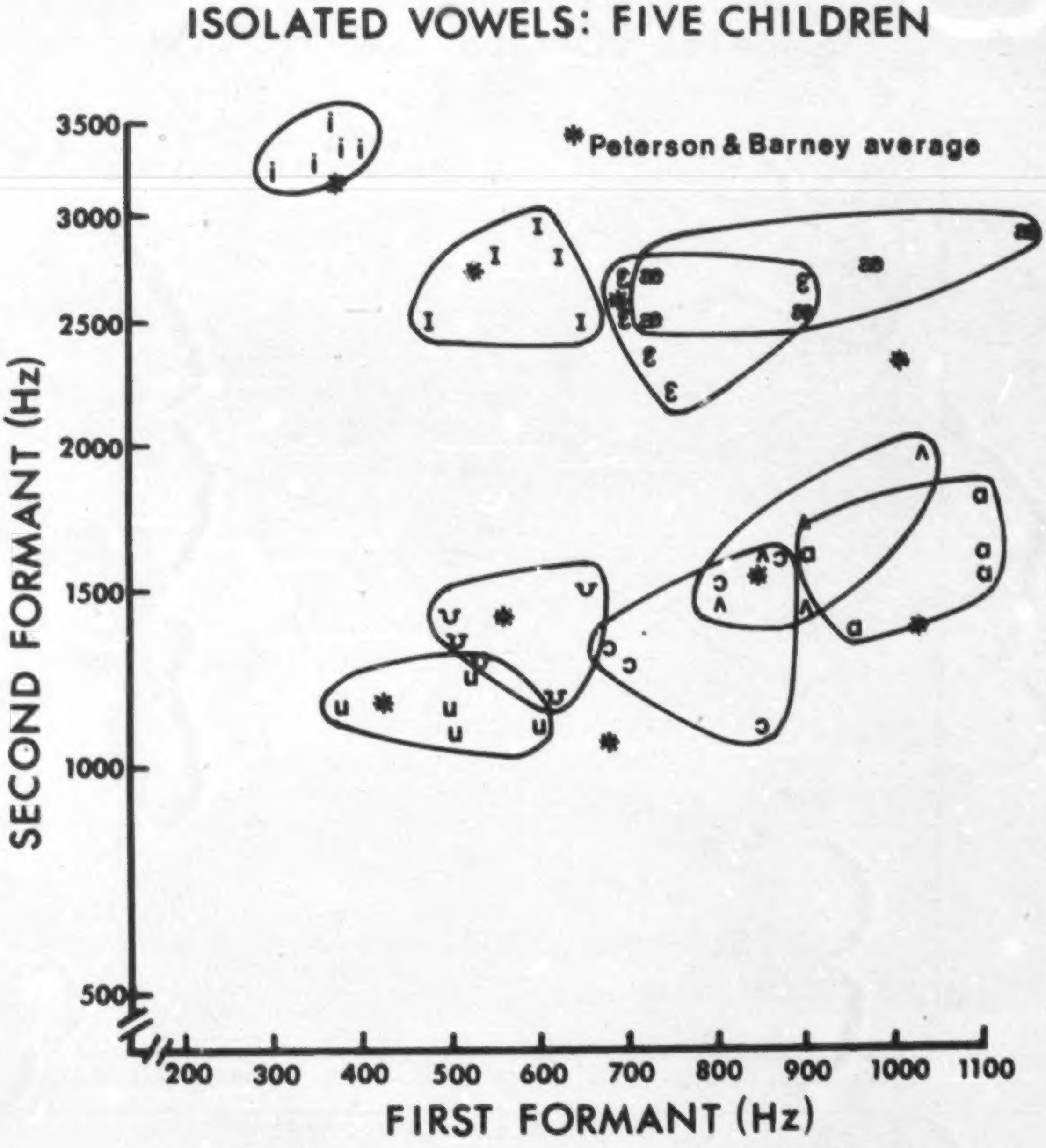

Figure 2: F_1/F_2 plots of five tokens of nine English vowels spoken by five children.

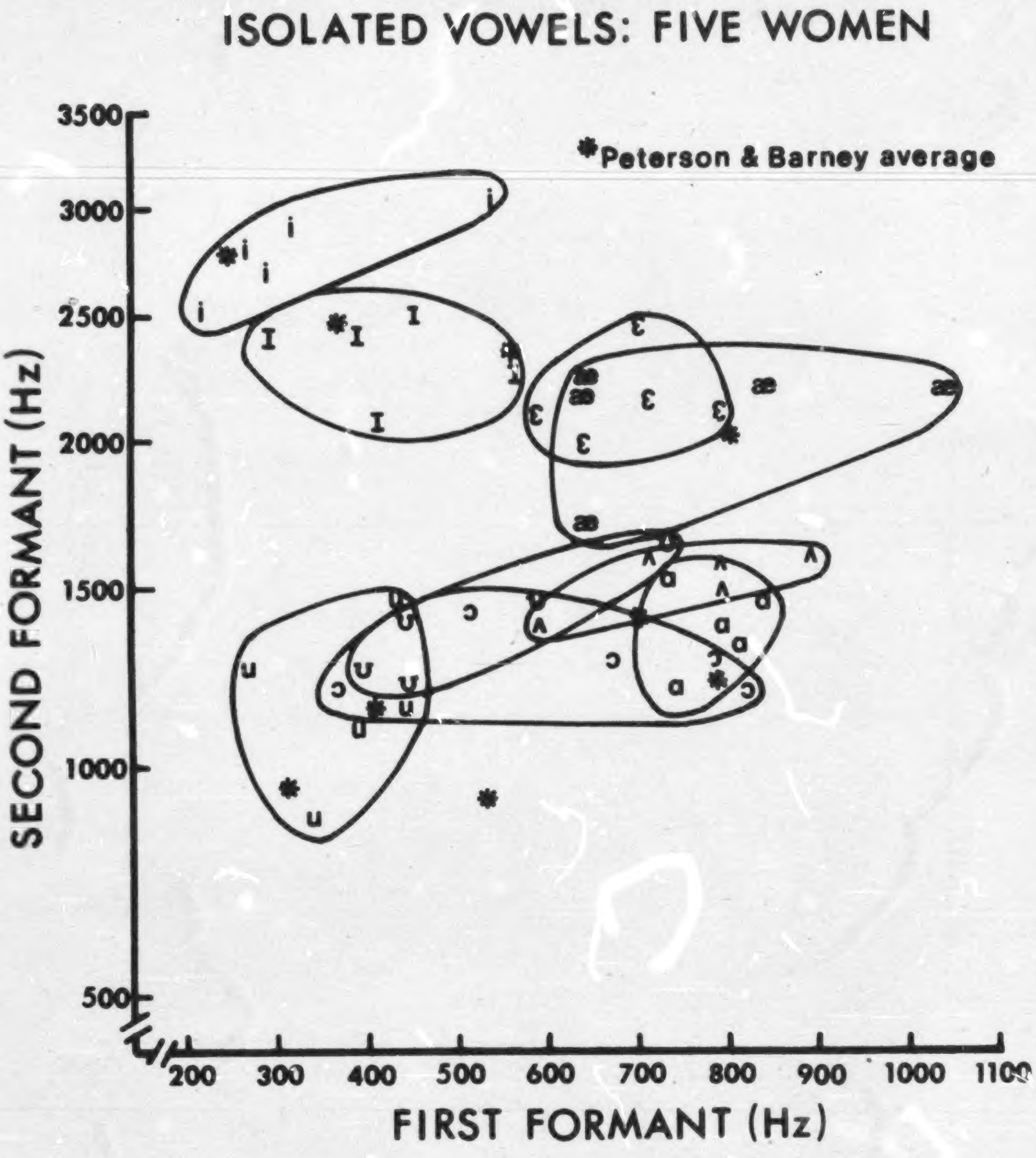

Figure 3: F_1/F_2 plots of five tokens of nine English vowels spoken by five adult females.

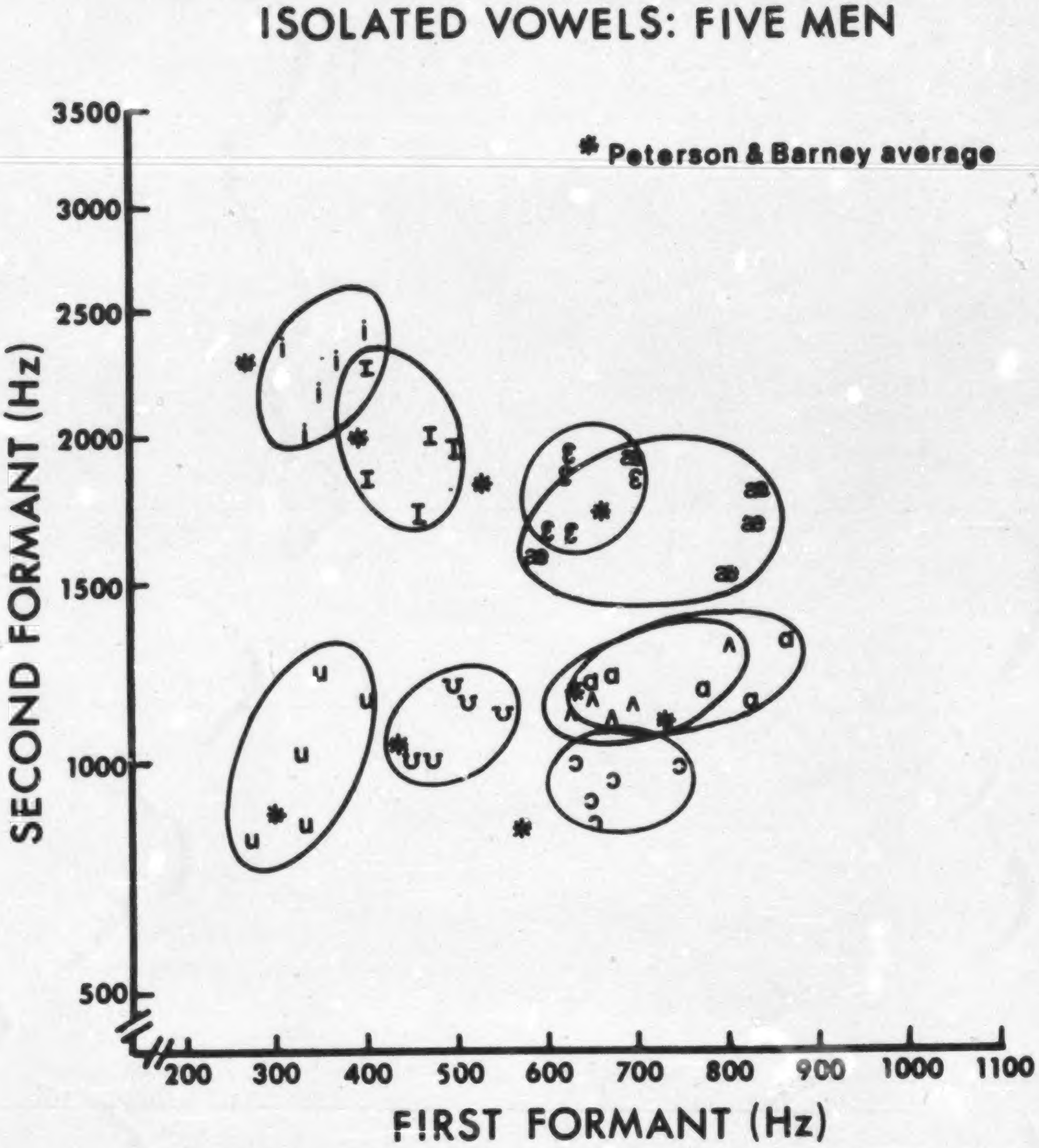

Figure 4: F_1/F_2 plots of five tokens of nine English vowels spoken by five adult males.

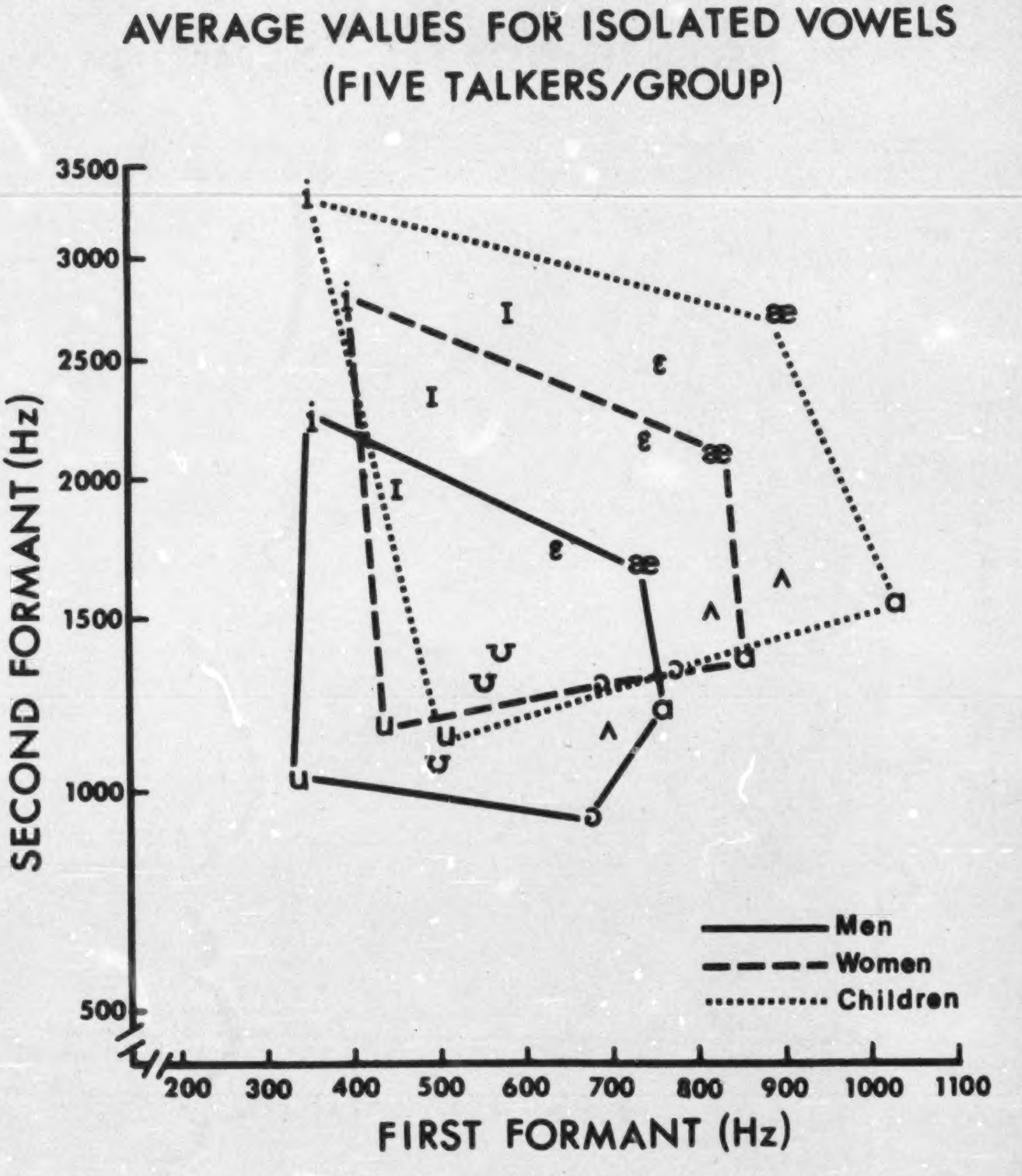

Figure 5: Mean F_1/F_2 points for five tokens of nine English vowels averaged separately for men, women, and children.

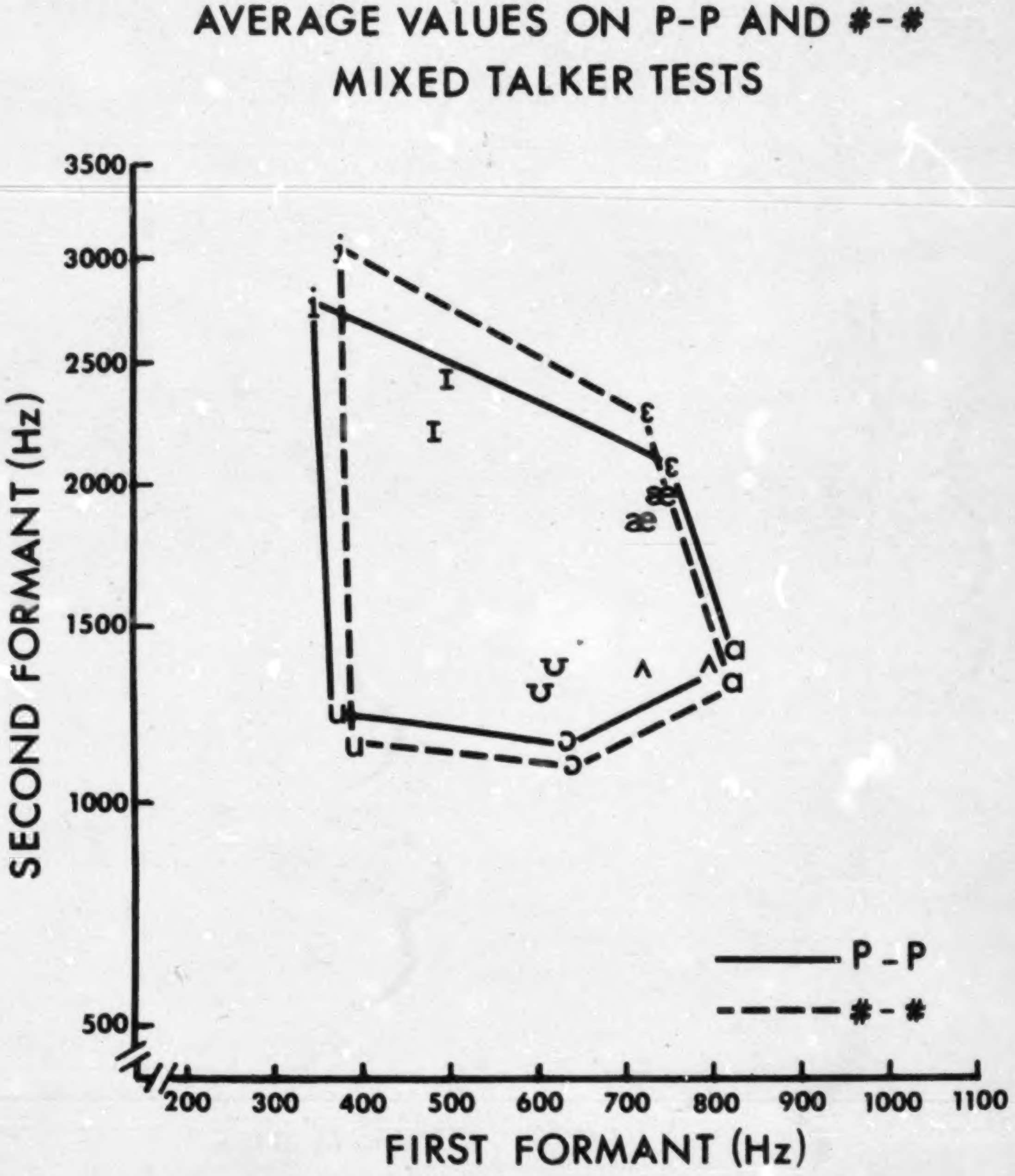

Figure 6: Mean F_1/F_2 points for five tokens of nine English vowels in /p-p/ environment and in null (#-#) environment spoken by 15 talkers, randomly mixed.

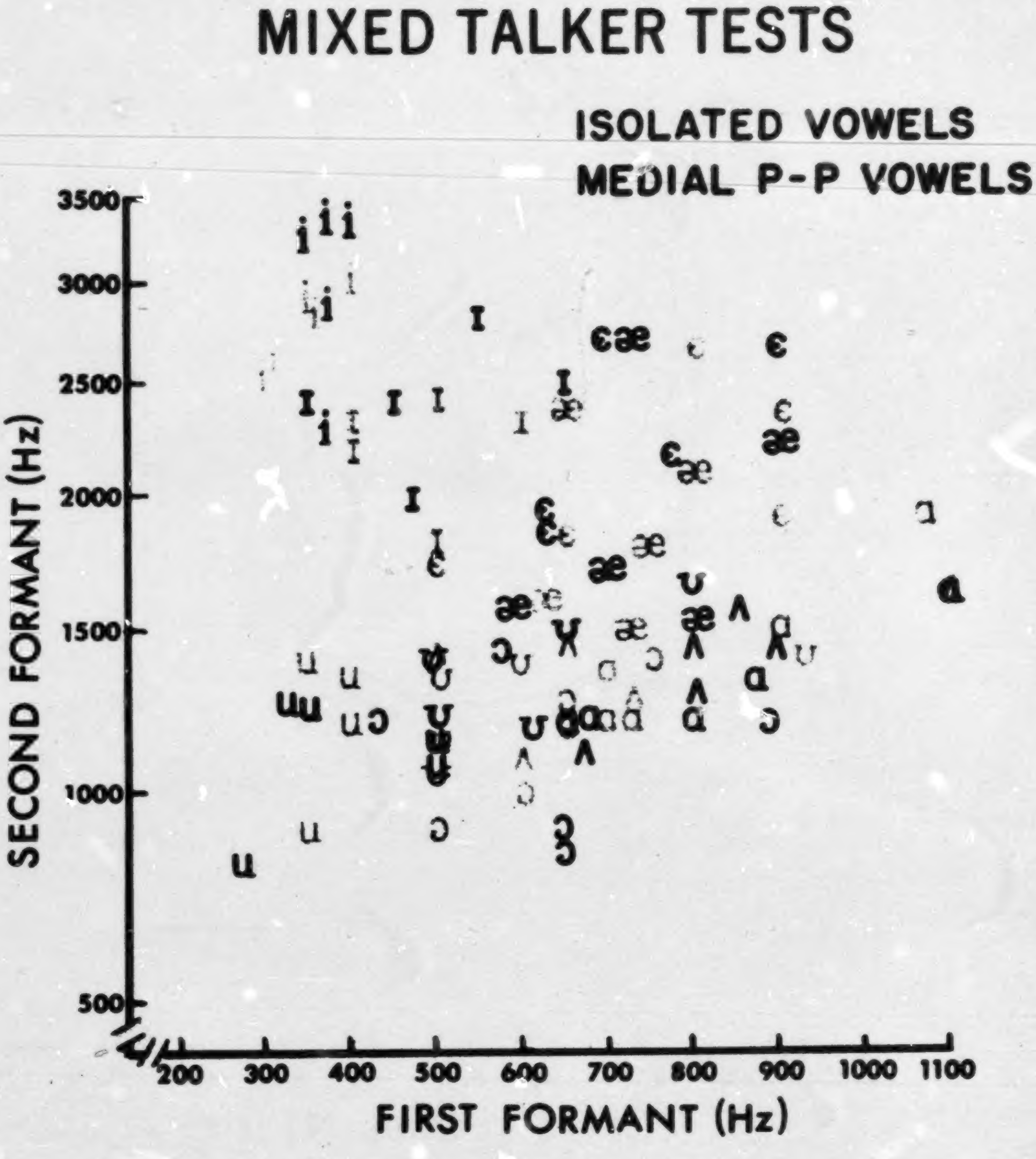

Figure 7: F_1/F_2 plot of 45 vowel tokens spoken in /p–p/ environment and in null (/#–#/) environment by 15 talkers.

increase in error rate for all isolated vowels. Although the durations of the latter were considerably longer than vowels in /p-p/ environment, the relative durations of the isolated vowels were the same as for vowels in /p-p/ position, except for /ʊ/, /ɛ/, and /i/, which were relatively longer in isolation than in context.

Having tentatively ruled out a theoretically uninteresting explanation of the difference in intelligibility between these two sets of vowels, we may ask what accounts, then, for the superior intelligibility of vowels in the stop environment. Surely this is puzzling, given the usual assumptions about the nature of the acoustic information that specifies vowel quality. An isolated, quasi steady-state utterance in which the formants attain appropriate targets ought to be an optimal signal for perception. Indeed, synthetic steady-state vowels based on these formant parameters are fairly intelligible to listeners. Moreover, in the domain of automatic speech recognition, some success has been achieved with a static model of the vowel. Gerstman (1968) devised an algorithm based on values of F_1 and F_2 derived from spectrographic measurements of center formant frequencies of /h-d/ syllables recorded from 76 talkers by Peterson and Barney (1952). Gerstman's algorithm sorted nine vowels in this set with only 2.5 percent error, less than was made by human listeners. From such a result, one might infer that target formant frequencies can in principle unambiguously specify the vowels of English as produced by a variety of talkers. It is but a short step to the conclusion that a human listener's strategy in identifying vowels is to extract the target formant values by means of something like a filter bank.

However, as we saw, this conception of the vowel cannot be reconciled easily with certain facts of perception. Since vowels in isolation were poor signals from the perceiver's standpoint, even though talkers adopted appropriate targets (differing little from /p-p/ targets), we can conclude that target frequencies do not adequately specify a vowel. Cues that we ordinarily regard as consonantal must contribute to the perception of the vowel. We suspect that much vowel information is contained in formant transitions, as Lindblom and Studdert-Kennedy (1967) suggested some time ago. In an analysis of perceptual adjustments for differences in stress and speaking rate, these investigators found that vowel identifications varied with direction and rate of transitions even when the formant frequency values at syllable centers were held constant. A case for the importance of formant transitions in vowel perception might also be made from considerations, such as those raised by Liberman (1970), of the extent of variation in formant contours conditioned by phonetic context. In any case, we are planning experiments to test the hypothesis directly by studying the effects of different consonantal environments, with and without transitions, on the perception of a coarticulated vowel.

Whatever the nature of the contribution consonantal environment makes to the identification of a vowel, the data point to the general conclusion that no single temporal cross section of a syllable conveys as much vowel information to a perceiver as is given in the dynamic contour of the formants. Thus it would seem that the definition of a vowel, from the standpoint of perception, ought to include a specification of how the relevant acoustic parameters change over time. If this conclusion is correct, then the specification of the relation between sound and percept presents the same problems for vowels as for consonants.

REFERENCES

Fairbanks, G. and P. Grubb. (1961) A psychophysical investigation of vowel formants. *J. Speech Hearing Res. 4*, 203-219.

Fujimura, O. and K. Ochiai. (1963) Vowel identification and phonetic contexts. *J. Acoust. Soc. Amer.*, *Suppl. 35*, 1889(A).

Gerstman, L. J. (1968) Classification of self-normalized vowels. *IEEE Trans. Audio Electroacoust. 16*, 78-80.

Lehiste, I. and D. Meltzer. (1973) Vowel and speaker identification in natural and synthetic speech. *Lang. Speech 16*, 356-364.

Liberman, A. M. (1970) The grammars of speech and language. *Cog. Psychol. 1*, 301-323.

Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. *J. Acoust. Soc. Amer. 35*, 1773-1781.

Lindblom, B. E. F. and M. Studdert-Kennedy. (1967) On the role of formant transitions in vowel recognition. *J. Acoust. Soc. Amer. 42*, 830-843.

Peterson, G. E. and H. L. Barney. (1952) Control methods used in the study of the vowels. *J. Acoust. Soc. Amer. 24*, 175-184.

Stevens, K. N. and A. S. House. (1963) Perturbation of vowel articulations by consonantal context: An acoustical study. *J. Speech Hearing Res. 6*, 111-128.

Strange, W., R. Verbrugge, and D. Shankweiler. (1974) Consonant environment specifies vowel identity. *J. Acoust. Soc. Amer.*, *Suppl. 55*, S54(A). [Full text in *Haskins Laboratories Status Report on Speech Research SR-37/38* (1974), 209-216.]

Some Acoustic Measures of Anticipatory and Carryover Coarticulation*

Fredericka Bell-Berti[+] and Katherine S. Harris[++]

ABSTRACT

Knowing the anticipatory and carryover limits of coarticulation will allow us to specify the constant features of a set of speech sounds. Most earlier work has reported on observations of the articulator movement or acoustic result of coarticulation. The present study is an attempt to define the extent and effects of coarticulation on the speech acoustic signal. Preliminary results suggest that carryover effects are more extensive than anticipatory effects.

INTRODUCTION

The nature and extent of coarticulation is of central interest to theories of speech production. Previous work on this problem, for several languages, has shown that anticipatory (or right-to-left) effects are either equal to, or greater in extent than, carryover (or left-to-right) effects, and that anticipatory effects may be different in cause from carryover effects (Daniloff and Hammarberg, 1973).

More specifically, Kozhevnikov and Chistovich (1965) and Daniloff and Moll (1968) have found anticipatory effects to extend over as many as three phoneme segments and across syllable boundaries. These effects have been explained as the reorganization of motor patterns for speech segments. Carryover effects, on the other hand, have often been attributed to mechanical inertia or articulator "sluggishness" (Lindblom, 1963; Stevens and House, 1963; Henke, 1966; Stevens, House, and Paul, 1966; MacNeilage, 1970), although these effects are now sometimes considered to be a deliberate reorganization of speech segments in the same way anticipation is a deliberate reorganization (MacNeilage and deClerk, 1969; Sussman, MacNeilage, and Hanson, 1973; Ushijima and Hirose, 1974).

---

*A version of this paper, under the title "Coarticulation in VCV and CVC Utterances: Some EMG Data," was presented at the 89th meeting of the Acoustical Society of America, Austin, Tex., 7-11 April 1975.

[+] Also Montclair State College, Upper Montclair, N. J.

[++] Also The Graduate School and University Center of the City University of New York.

[HASKINS LABORATORIES: Status Report on Speech Research SR-42/43 (1975)]

In spite of the central position of coarticulation rules in a general theory of speech production, there are very few descriptive data on the relative magnitude of anticipatory and carryover coarticulation effects at an acoustic level. The experiment to be described was an attempt to fill this gap in our knowledge.

PROCEDURE

The utterance set contained 18 three-syllable nonsense words, consisting of a stressed consonant-vowel-consonant (CVC) preceded by [pə] and followed by [əp]. The vowel in the stressed syllable was either /i, a, or u/, and the consonants were /p, t, k/. All combinations of consonants and vowel were used, except the symmetric ones, such as /pəpikəp/, /pətupəp/, and /pəkatəp/. The utterances were spoken within a carrier phrase, "Say __________ now," at a conversational rate of speech.

Acoustic recordings were obtained, from one speaker of American English, of 18 repetitions of each of the 18 utterance types.

The audio signal was sampled through the Haskins Laboratories pulse-code-modulation (PCM) and Spectrum Analyzing Systems, the former for editing, the latter for generating spectrum data. After software filtering (and thresholding), hard copies of computer-generated spectrograms were obtained and formant measurements made offline.

Since second-formant position is extremely sensitive to back-to-front tongue position and lip-rounding--that is, front cavity length--F_2 measurements were made at seven points in each repetition of each utterance type. Averages of 15 to 18 measurements for each sample point were obtained, corresponding to those tokens included in the EMG average for each utterance type. Schematic spectrograms of F_2 were generated from these averages.

The measurement points were

1. One point in $ə_1$;

2. The beginning, middle, and end points of the stressed vowel;

3. The beginning, middle, and end points of $ə_2$.

No attempt was made to account for durational variation, since the sample time represented by each data point in the spectrogram is 12.8 msec; hence, the time scale is too crude for detailed measurements.

RESULTS

The results are summarized in Figures 1 and 2; the first shows the 18 utterances plotted with the first consonant held constant; the second shows the same data with the second consonant held constant.

Beginning with the stressed vowel, we see that the initial point is determined by the preceding consonant, and the end point is affected by the following consonant. The effects of the terminal consonant on the midpoint of the stressed

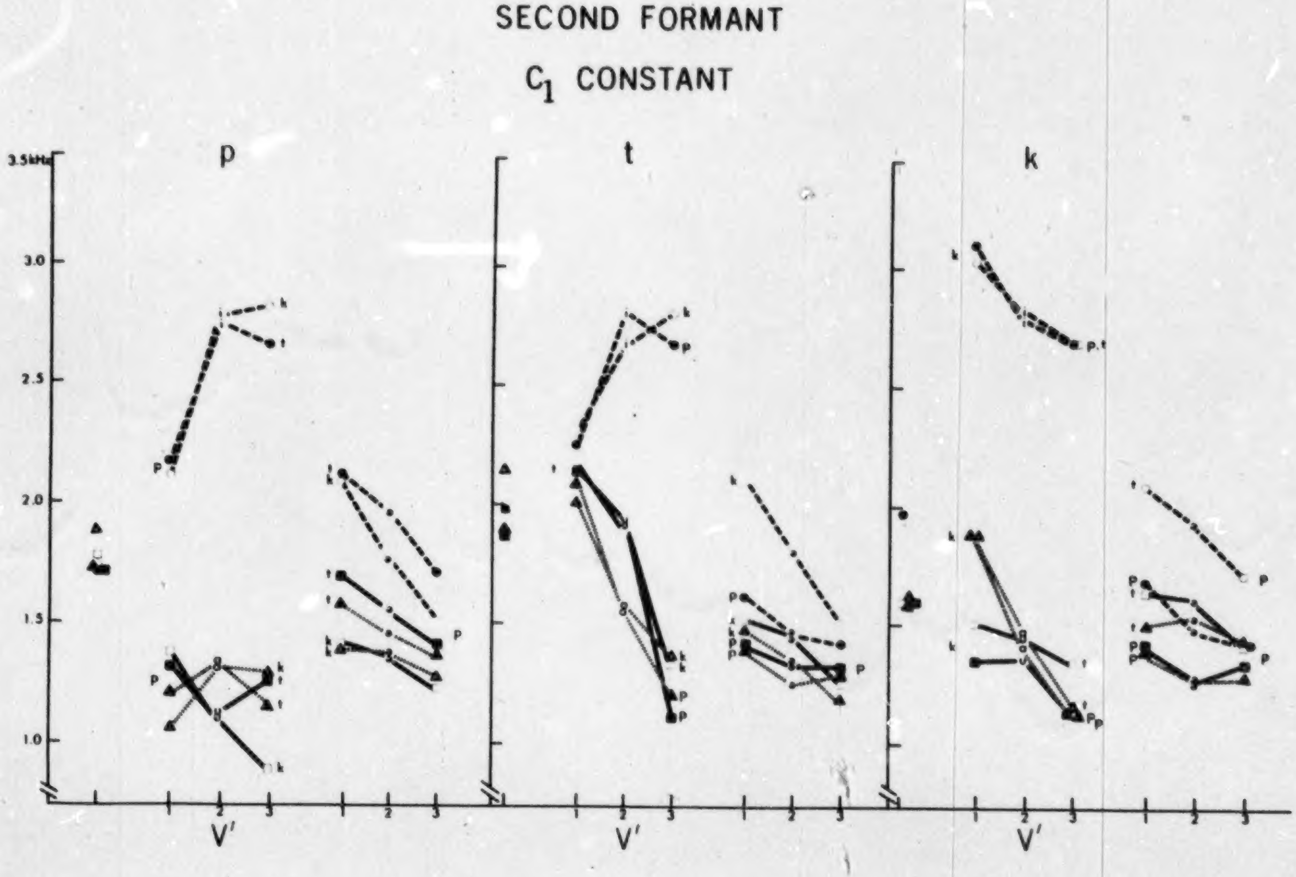

Figure 1: All measurements are of F_2. The single data points at the left of each section are $ə_1$ measurements. The three-point plots above V' in each section are stressed vowel measurements. The three-point plots to the right of each section are $ə_2$ measurements. The left section is F_2 averages for utterances whose stressed syllables begin with /p/; the middle section, with /t/; the right section, with /k/.

FIGURE 1

SECOND FORMANT

C_2 CONSTANT

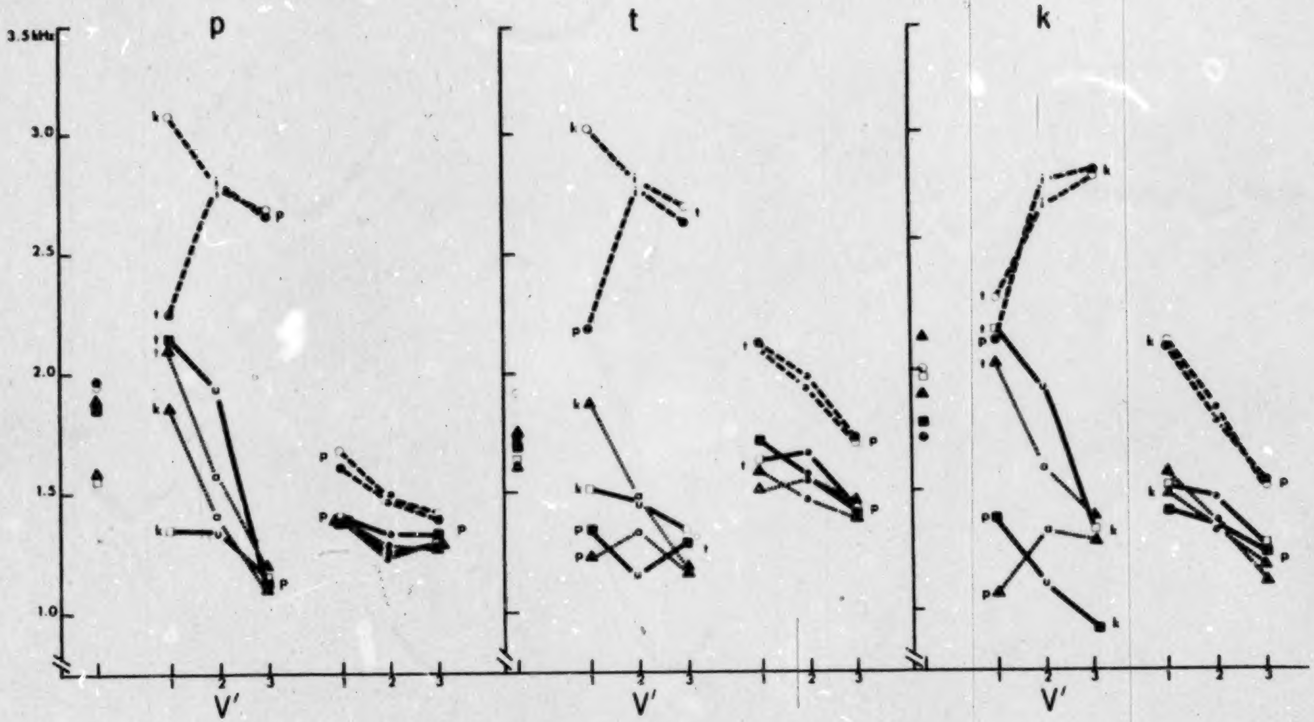

Figure 2: All measurements are of F_2. The single data points at the left of each section are $ə_1$ measurements. The three-point plots above V' in each section are stressed vowel measurements. The three-point plots to the right of each section are $ə_2$ measurements. The left section is F_2 averages for utterances whose stressed syllables end with /p/; the middle section, with /t/; the right section, with /k/.

FIGURE 2

vowel are not as large as those of the initial consonant. The midvowel position is more often determined by the initial consonant. In other words, the carryover effect of the first consonant on the stressed vowel is larger than the anticipatory effect of the second.

We can also examine the relative magnitudes of anticipatory and carryover effects by looking at the effects of environment on the initial and terminal schwa vowels. One-step effects are seen in both directions--the initial schwa is affected by the following consonant, while the second schwa is affected by the preceding consonant. However, when we turn to the two-step effects, we find that the initial schwa is not affected by the following vowel, while the same vowel does change the value of the following schwa. In general, then, at the acoustic level, carryover effects are larger than anticipatory effects. It is this asymmetry of effect that must be accounted for at an articulatory level.

REFERENCES

Daniloff, R. and R. Hammarberg. (1973) On defining coarticulation. J. Phonetics 2, 239-248.

Daniloff, R. and K. Moll. (1968) Coarticulation of lip rounding. J. Speech Hearing Res. 11, 707-721.

Henke, W. (1966) Dynamic articulatory model of speech production using computer simulation. Unpublished Ph.D. dissertation, Massachusetts Institute of Technology.

Kozhevnikov, V. A. and L. A. Chistovich. (1965) (in translation) Speech, Articulation and Perception (Washington, D.C.: Joint Publications Research Service, U. S. Department of Commerce No. 30).

Lindblom, B. E. F. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. 35, 1773-1781.

MacNeilage, P. F. (1970) The motor control of serial ordering of speech. Psychol. Rev. 77, 182-196.

MacNeilage, P. F. and J. L. deClerk. (1969) On the motor control of coarticulation in CVC monosyllables. J. Acoust. Soc. Amer. 45, 1217-1233.

Stevens, K. N. and A. S. House. (1963) Perturbation of vowel articulation by consonantal context: An acoustical study. J. Speech Hearing Res. 6, 111-128.

Stevens, K. N., A. S. House, and A. P. Paul. (1966) Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation. J. Acoust. Soc. Amer. 40, 123-132.

Sussman, H. M., P. F. MacNeilage, and R. J. Hanson. (1973) Labial and mandibular dynamics during the production of bilabial stop consonants. J. Speech Hearing Res. 16, 397-420.

Ushijima, T. and H. Hirose. (1974) Electromyographic study of the velum during speech. J. Phonetics 2, 315-326.

II. PUBLICATIONS AND REPORTS

III. APPENDIX

PUBLICATIONS AND REPORTS*

Abramson, A. S. (in press) The tones of Central Thai: Some perceptual experiments. In _Studies in Tai Linguistics in Honor of William S. Gedney_, ed. by Jimmy G. Harris and James R. Chamberlain. (Bangkok: Central Institute of English Language), pp. 1-16.

Ingemann, F. and P. Mermelstein. (1975) Speech recognition through spectrogram matching. _Journal of the Acoustical Society of America_ _57_, 253-255.

Kuhn, G. M. (1975) On the front cavity reasonance and its possible role in speech perception. _Journal of the Acoustical Society of America_ _58_, 428-433.

Lisker, L. (1975) Is it VOT or a first-formant transition detector? _Journal of the Acoustical Society of America_ _57_, 1547-1551.

---

*Henceforth only articles or chapters that have already been published and do not appear in the present _SR_ will be cited in this section.

APPENDIX

DDC (Defense Documentation Center) and ERIC (Educational Resources Information Center) numbers:

SR-21/22 to SR-41

Status Report		DDC	ERIC
SR-21/22	January - June 1970	AD 719382	ED-044-679
SR-23	July - September 1970	AD 723586	ED-052-654
SR-24	October - December 1970	AD 727616	ED-052-653
SR-25/26	January - June 1971	AD 730013	ED-056-560
SR-27	July - September 1971	AD 749339	ED-071-533
SR-28	October- December 1971	AD 742140	ED-061-837
SR-29/30	January - June 1972	AD 750001	ED-071-484
SR-31/32	July - December 1972	AD 757954	ED-077-285
SR-33	January - March 1973	AD 762373	ED-081-263
SR-34	April - June 1973	AD 766178	ED-081-295
SR-35/36	July - December 1973	AD 774799	ED-094-444
SR-37/38	January - June 1974	AD 783548	ED-094-445
SR-39/40	July - December 1974	AD A007342	ED-102-633
SR-41	January - March 1975	AD A103325	ED-109-722

AD numbers may be ordered from:
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22151

ED numbers may be ordered from:
ERIC Document Reproduction Service
Computer Microfilm International Corp. (CMIC)
P.O. Box 190
Arlington, Virginia 22210